A Complex Network Approach to Phylogenetic Trees: From Genes to the Tree of Life

TESI DOCTORAL

E. Alejandro Herrada

Directors:
Prof. Emilio Hernández- García
Dr. Víctor M. Eguíluz
Prof. Carlos M. Duarte

Ponent:
Prof. José A. Castro Ocón

Presentada al Departament de Biologia
Universitat de les Illes Balears
2010

A Complex Network Approach to Phylogenetic Trees: From Genes to the Tree of Life

E. Alejandro Herrada
Tesi presentada al Departament de Biologia de la Universitat de les Illes Balears

PhD Thesis

Directors: Prof. Emilio Hernández-García, Dr. Víctor M. Eguíluz and Prof. Carlos M. Duarte

This document was typeset with LaTeX 2_ε

Tesi doctoral presentada per E. Alejandro Herrada per optar al títol de Doctor, en el Programa de Biologia del Departament de Biologia de la Universitat de les Illes Balears, realitzada a l'IFISC sota la direcció de Emilio Hernández-García, Professor de Investigació del CSIC (Consejo Superior de Investigaciones Científicas), Víctor M. Eguíluz, Científic Titular del CSIC i Carlos M. Duarte, Professor de Investigació del CSIC, i amb José A. Castro Ocón, Catedràtic d'Universitat, com a ponent.

Vist i plau
Directors de la tesi

Prof. Emilio Hernández-García	Dr. Víctor M. Eguíluz	Prof. Carlos M. Duarte

Ponent	Doctorant
Prof. José A. Castro Ocón	E. Alejandro Herrada

Palma, 02 de novembre de 2010

A Lu y Jutta

"All animal life has the right to be respected." — Art. 2 of the Universal Declaration of Animal Rights (UNESCO, 1990)

Acknowledgments

Quisiera agradecer a mis directores, el Prof. Emilio Hernández García, el Dr. Víctor M. Eguíluz y el Prof. Carlos M. Duarte por haberme dado la oportunidad de realizar esta tesis, así como por el tiempo dedicado y las discusiones mantenidas. En especial, gracias, Emilio, por tu generosidad y humildad. De ti he aprendido muchas más cosas que las relacionadas con el análisis de árboles filogenéticos. Muchas gracias.

Adrián, gracias por toda la ayuda que me has proporcionado a lo largo de estos años. Sabes que sin ella esta tesis difícilmente hubiera avanzado después del segundo año de investigación. Fuiste mi cuarto director, y lo sabes.

Gracias, Maxi San Miguel, por haberme abierto la puerta aquella mañana de septiembre de 2005. Nadie habría imaginado que esa visita daría lugar a toda esta aventura.

Gracias a Pepe Castro por su disponibilidad y amabilidad.

I would also like to thank Prof. Kathleen Marchal and her people for having accepted me, and for their hospitality during my stay in KU Leuven.

Gracias, Joan, Carlos y Mar, por todo vuestro tiempo y *feedback*.

Thanks are also due to the EDEN people, specially to Sophie and Konstantin for the numerous, fruitful discussions.

Gracias a los contribuyentes europeos, en especial a los de las Illes Balears, por haber financiado esta tesis a través de una beca predoctoral del Govern de les Illes Balears, así como por medio de los proyectos europeos THRESHOLDS y EDEN.

Gracias a Rubén, Edu y M. Antònia, por cuidar de Nuredduna y de todos nosotros.

Gracias a aquellos con los que he compartido alguno de estos años en la EFE o en el sótano, especialmente a mis hermanos de tesis: Xavi, por tu compromiso, Adrián, por tu lealtad, Juan Carlos, por tus abrazos, Niko, por tu irreverencia, Leo, por tu alegría, Murat, por tu honestidad, Juan, por tu bondad.

Gracias a la gente que me acompañó durante la carrera: Israel, Muro, Piti, Rocío, Adrián, Begoña. Gracias a Anadón, "Primi", Zapata y Marcos, fuentes de motivación. Del mismo modo, me gustaría agradecer a la gente que me acompañó en mis primeros pasos en el mundo de la investigación: Luz, Antía, Pili, Ruth, Olga, Susana, Teresa, Eduardo, Rubén, Carmen Carneiro, Ramón Castro y Ramón Ríos. Gracias a Antonio Piñeiro y Fernando Domínguez, por haber entendido mi necesidad de marchar.

Gracias a Don Lorenzo y a "El Ficus", por haberme enseñado a "echar reaños". Y a Fábregas, por haberme enseñado la utilidad del descaro comedido.

Gracias, familia de S'Esgleieta y de Mar del Plata. Siempre nos quedará Garmisch.

Gracias, Fer, por dejarme participar de tu novela.

Gracias a todos los que han ido apareciendo entre filosofías, feminismos y ecologismos. Andrea...

Gracias a Rafa por haberme acercado la biología desde Nano, y a Jose por haberme enganchado a esa droga que es el conocimiento alegre.

Gracias a los de Pontevedra, por esa vida que dais.

Gracias a mi padre. Perdón por todo lo sufrido.

Gracias, Elsa, porque hay cosas que cambian ;-).

Gracias, Jutta, porque nunca dejamos de ser camaradas. Ajax, tú también estás aquí.

Y a Lucía, por lo enseñado, lo aprendido, lo compartido...

Contents

Preface

The increasing interest during the last century in the study and comprehension of the evolutionary processes that govern biodiversity, as well as the huge expansion that the complex network approach has undergone in the last decade, has motivated us to address the interrelation of both scientific fields. In that sense, the main goal of this thesis is the application of the complex network theory to the inference of evolutionary patterns through the topological characterization of evolutionary trees.

In Chapter 1 we will introduce some of the most relevant concepts derived from evolutionary biology and phylogenetics, as well as a short overview about the application of complex network theory to evolutionary biology, with a short description of some of the most outstanding applications of the complex network theory to the study of biological evolution, and with a summary of some basic concepts derived from the complex network theory useful for the analysis of evolutionary trees.

Chapter 2 presents the theoretical foundation of this study, i.e. it offers a review of some of the most used measures for the characterization of the topological properties of the evolutionary trees. We

will propose as well the application of the depth scaling analysis, a specific complex network approach based on the allometric scaling relationships between size and shape of the tree-like networks, for the topological characterization of evolutionary trees. This theoretical chapter will be completed with the description of two of the most relevant evolutionary models.

In Chapter 3 we apply the depth scaling approach to a comparative analysis between micro- and macroevolutionary phylogenies from organisms distributed all over the Tree of Life. The lists of works used for the compilation of a dataset of intraspecific and interspecific phylogenies are included in Appendix A. Moreover, in Appendix B, we include a short analysis about the effect of the outgroups over the allometric scaling of the phylogenetic trees.

In Chapter 4 we extend the comparative analysis carried out in Chapter 3 to the molecular level, comparing gene versus organism evolutionary trees. With the aim of going deeper in the understanding of the evolutionary mechanisms that shape the diversification of gene families, in Appendix C, we try to depict to what extent speciation and gene duplication events contribute to protein family diversification.

In order to propose an alternative evolutionary mechanism that explains the results obtained in Chapters 3 and 4, in Chapter 5 we describe an evolutionary model based on the biological concept of *evolvability*, referred to the ability of a new species or a protein to evolve. Besides, in Appendix E we analyze the effect of refractory period between consecutive diversification events and the effect of mass extinction events over the depth scaling behavior of the evolvability model. Furthermore, in Appendix D, we propose the *activity model*, an evolutionary model characterized by depicting a non-ERM depth scaling.

In Chapter 6 we extend the depth scaling approach for the characterization of the effects of the rank-based and rank-free taxonomic criteria over the topological properties of the evolutionary trees, and

in Appendix F we extrapolate this comparative analysis between rank-based and rank-free taxonomic criteria to language evolutionary trees.

In Chapter 7 we take a first step toward the characterization of the branch length distribution all over the Tree of Life, and in Appendix G we propose a set of measures for the characterization of the depth scaling taking into account the branch length of the evolutionary trees.

Finally, in Chapter 8 we summarize the results obtained, and give some concluding remarks.

We also include, in Appendix H, the Python codes used for the computation of the depth scaling analysis, for the conversion of tree files from Newick format to columns format, as well as the Python code used for the simulation of the evolvability model. In addition, Appendix I includes the detailed list of publications derived from this thesis.

The original research of this thesis is contained mainly in Chapters 3 to 8, and in the Appendices, although some of the theoretical foundations in Chapters 1 and 2 also contain original material.

The datasets analyzed in this thesis have been compiled in the URL `http://ifisc.uib-csic.es/~alejandro/phylotreedata/`.

Chapter 1

Introduction

1.1 Biological evolution at a glance: From molecules to organisms

A word closely related to evolution is *change*. It is said that a system evolves when this system undergoes a change over time. Thus, biological evolution refers to the accumulation of inheritable changes (*mutations*) in a biological system over time. The inheritability of mutations is given by the fact that they occur in the nucleic acid molecule that constitutes the *genome*, the inheritable material, of the organism.[1] Together with mutations, three other mechanisms, i.e. migration, genetic drift and selection, constitute the four main forces responsible for biological evolution (Freeman and Herron, 2001). These forces can take place at three main evolutionary levels:

[1] In multicellular organisms, in order to guarantee the inheritability of the mutations that take place at the genome of the organism, those mutations have to take place at the genome of the germ cells, since they are the cells in charge of giving rise to the next generation of organisms and therefore, they are responsible for passing the new mutations on to the next generation.

- Evolution at the individual level.
- Evolution at the population level (microevolution).
- Evolution at the species level (macroevolution).

Evolution at the individual level

The label *evolution at the individual level* is used to refer to those mechanisms responsible for the inheritable changes that take place in a single organism. Those mechanisms take place at a genomic level, originating changes at single nucleotides (*small-scale mutations*) or changes that affect sequences of nucleotides (*large-scale mutations*)

[2]See Guthrie (1962).
[3]See Aristóteles (1994).
[4]See von Linné (1758).
[5]See Burnett (1974).
[6]See Darwin (1794-1796).
[7]See Winchester (2001); Cuvier and Brongniart (1822).
[8]See Lamarck (1809).
[9]See Wells (1818).
[10]See Hitchcock (1840).
[11]See Darwin and Wallace (1858).
[12]See Darwin (1859).
[13]See Mendel (1865).
[14]See Haeckel (1866).
[15]See Weismann (1892).
[16]See Wallace (1889).
[17]See Nuttall (1904).
[18]See Fisher (1930); Haldane (1932); Wright (1931, 1932).
[19]See Dobzhansky (1937).
[20]See Avery et al. (1944).
[21]See Franklin (1952).
[22]See Watson and Crick (1953).
[23]See Margoliash (1963).
[24]See Kimura (1968).
[25]See Eldredge and Gould (1972).

610-546 BC	Anaximander: First animals lived in water and originated the land animals.[2]
384-322 BC	Aristotle: First classification of the living forms.[3]
1735	C. von Linné: Rank-based classification of living organisms.[4]
1773-1792	J. Burnett: Human being had descended from primates.[5]
1794-1796	E. Darwin: Warm-blooded animals arose from one living filament.[6]
1790-1811	W. Smith, G. Cuvier & A. Brogniart: Principle of faunal succession.[7]
1809	J.-B. Lamarck: Theory of transmutation of species, based on increasing complexity and adaptation. Evolutionary tree of animals.[8]
1813	W.C. Wells: Assigned a role to the natural selection in the human evolution.[9]
1840	E. Hitchcock: Evolutionary trees, based on paleontology data, of plants and animals, without connection between them.[10]
1858	C. Darwin & A.R. Wallace: Natural selection is the basic mechanism of evolution.[11]
1859	C. Darwin: Theory of evolution based on natural selection. A single Tree of Life, with a common ancestor, as a sketch of the evolution.[12]
1865	G. Mendel: Theory of particulate inheritance.[13]
1866	E. Haeckel: First labeled Tree of Life.[14]
1883	A. Weismann: Germ-plasm theory. First neo-darwinist work.[15]
1889	A.R. Wallace: One of the first proponents of neo-darwinism.[16]
1904	G.H.F. Nuttall: Phylogenetic relationships among different groups of animals through conducted precipitin tests of serum protein.[17]
1920-1930s	R.A. Fisher, J.B.S. Haldane & S. Wright: Foundation of population genetics.[18]
1937	T. Dobzhansky: Publication of the major work of the modern evolutionary synthesis.[19]
1944	O. Avery: Identification of the DNA as the genetic material.[20]
1952	R. Franklin: X-ray diffraction image of the DNA molecule.[21]
1953	J.D. Watson & F. Crick: Double-helix model of the DNA structure.[22]
1963	E. Margoliash: Cytochrome c phylogeny for horse and other species.[23]
1968	M. Kimura: Neutral theory of evolution.[24]
1972	N. Eldredge & S.J. Gould: Punctuated equilibrium theory.[25]

Table 1.1: Some of the main events in the history of evolutionary thought.

(Freeman and Herron, 2001; Tamarin, 1996; Griffiths et al., 2000; Freeman and Herron, 2001).

Examples of small-scale mutations are:

- *Point mutation (silent, missense, nonsense)*: Substitution of a single nucleotide by another one.
- *Insertion*: Addition of one or more extra nucleotides in the DNA sequence.
- *Deletion*. Elimination of one or more nucleotides from the DNA sequence.

The most common large-scale mutation processes are:

- *Amplification (or gene duplication)*: Multiplication of a chromosomal region.
- *Insertion*: Addition of an extra chromosomal region.
- *Deletion*: Loss of a chromosomal region.
- *Chromosomal inversion*: 180 degrees rotation of a chromosomal segment.
- *Chromosomal recombination*: A chromosomal region exchange between two homologous chromosomes.
- *Chromosomal translocation*: A chromosomal region exchange between two nonhomologous chromosomes.
- *Chromosomal transpositions*: A chromosomal region relocation to a different position in the genome.
- *Euploidy*: The cell or the organism changes to an integer multiple of the haploid number of chromosomes.

Evolution at the population level: Microevolution

Evolution at the population level is said to take place when population dynamic events give rise to a change in the genetic pool of the population. Before describing the different mechanisms that give rise to evolution at population level, we are going to describe briefly the basic behavior of a non-evolving population. This scenario was described, independently, by the mathematician G. H. Hardy and by the physician W. Weinberg. Both postulated a law, known as the *Hardy-Weinberg principle*, that relates the allele and genotype frequencies in a diploid population with sexual reproduction (Hardy, 1908; Weinberg, 1908). For a diploid population with sexual reproduction, random mating, infinitely large population size, no mutation, no migration, and without any selection pressure, they established the following statements (Tamarin, 1996; Griffiths et al., 2000; Freeman and Herron, 2001; Halliburton, 2004):

1. *Equilibrium of the allele frequencies*. The allele frequencies for an autosomal locus do not change from one generation to the next.

2. *Equilibrium of the genotype frequencies*. The genotype frequencies of the population are determined, in a predictable way, by the allele frequencies.

3. *Neutral equilibrium*. If the population is perturbed, the equilibrium will be restored in a single generation of random mating, but with the new allele frequencies.

Based on these holds, considering a single autosomal locus with two alleles, A and a, and their corresponding allele frequencies, p and q, the Hardy-Weinberg equilibrium distribution for the genotype

frequencies in a diploid organism with discrete, nonoverlapping generations, would be:[26]

$$\begin{array}{ccc} AA & Aa & aa \\ p^2 & 2pq & q^2 \end{array} .$$

The non-evolving scenario proposed by Hardy and Weinberg (random mating, infinitely large population size, no mutation, no migration and without any selection pressure) is far away from nature, and modifications of the Hardy-Weinberg equilibrium for each of the deviations of those assumptions were proposed. The effect of the four main evolutionary forces (mutation, migration, genetic drift and selection), over the allele frequencies inside a population, is quantified as follows (Tamarin, 1996; Griffiths et al., 2000; Freeman and Herron, 2001; Halliburton, 2004):

- *Mutations*: As defined at the beginning of this section, mutations are all those changes that occur in the genome of an organism. In order to understand how it interferes in the allele frequency, let us consider the simplest case, with a mutation rate for an allele A, μ, as the probability that a copy of allele A becomes allele a in a DNA replication event. If p_0 is the frequency of allele A, after n generations of mutations, the frequency of allele A, p_n, assuming no back mutations, will be (assuming μ small):
$$p_n = p_0 e^{-n\mu} .$$

- *Gene flow (Migrations)*: It is the exchange of alleles between populations. The effect of the gene flow is similar to the effect of the mutations in the sense that it changes the allele frequencies adding or eliminating alleles. If p_t is the frequency of an

[26] Multiple extensions of the Hardy-Weinberg equilibrium have been described, such as that for multiallelic loci or that for the case of various loci (Tamarin, 1996; Halliburton, 2004).

allele in the recipient population in generation t, P is the allele frequency in a donor population, and m is the proportion of the recipient population that consists of new migrants arrived in one generation from the donor population, then the gene frequency in the recipient population in the next generation, p_{t+1}, is the result of mixing $1 - m$ genes from the recipient with m genes from the donor population. Thus:

$$p_{t+1} = (1 - m)p_t + mP = p_t + m(P - p_t)$$

and

$$\Delta p = p_{t+1} - p_t = m(P - p_t) \, .$$

- *Genetic drift (Neutral evolution)*: It is the change in the allele frequency from one generation to the next one, given by the random sampling of the parents. Supposing p_t as the frequency of an allele in generation t, the expected value for the mean at $t + 1$ will be:

$$E(p_{t+1}) = p_t$$

and the variance:

$$V(p_{t+1}) = \frac{p_t(1 - p_t)}{2N} \, ,$$

where N is the population size.

Variance equation gives us an idea about the magnitude of allele frequency changes from one generation to the next. So, genetic drift is basically given by the finite population size effect. The smaller the population, the larger the change from one generation to the next. The long-term effect of the genetic drift is the decrease of the genetic variation within a population and the divergence between populations.

- *Selective evolution*: It corresponds to the different degree of survival or reproduction, on average, of different traits in a population. This different survival or reproduction leads to changes in frequencies of those genotypes, within a population. If we consider a population in Hardy-Weinberg equilibrium, and we break this equilibrium through a selective process given by differential survival probabilities for the three possibles genotypes (AA, Aa and aa): W_{AA}, W_{Aa}, W_{aa}, the genotype frequencies for the zygotes are:

$$\begin{array}{ccc} AA & Aa & aa \\ p^2 & 2pq & q^2 \end{array},$$

while the genotype frequency for the adults will be:

$$\begin{array}{ccc} AA & Aa & aa \\ p^2 W_{AA} & 2pqW_{Aa} & q^2 W_{aa} \end{array}.$$

The sum of all the frequencies after selection will be smaller than 1, thus we have to normalize by the *mean fitness* of the population, $\overline{W}$:

$$\overline{W} = p^2 W_{AA} + 2pqW_{Aa} + q^2 W_{aa}\,.$$

After normalizing:

$$\begin{array}{ccc} AA & Aa & aa \\ p^2 \dfrac{W_{AA}}{\overline{W}} & 2pq \dfrac{W_{Aa}}{\overline{W}} & q^2 \dfrac{W_{aa}}{\overline{W}} \end{array}.$$

From this information we can obtain the allele frequencies in the next generation. So, for example, for the allele A, the allele frequency in the next generation, p_{t+1}, would be:

$$p_{t+1} = AA + \frac{1}{2}Aa = p^2\frac{W_{AA}}{\overline{W}} + \frac{pqW_{Aa}}{\overline{W}} = p\frac{pW_{AA} + qW_{Aa}}{\overline{W}} ,$$

where, considering that $\overline{W}_A = pW_{AA} + qW_{Aa}$, the final new frequency is:

$$p_{t+1} = p\frac{\overline{W}_A}{\overline{W}} .$$

An alternative way of looking at the process of selection is solving for the change in allele frequency in one generation:

$$\Delta p = p_{t+1} - p = \frac{p\overline{W}_A}{\overline{W}} - p = \frac{p(\overline{W}_A - \overline{W})}{\overline{W}} .$$

Taking into account that $\overline{W}$ is the average of the allele fitnesses $\overline{W}_A$ and $\overline{W}_a$:

$$\overline{W} = p\overline{W}_A + q\overline{W}_a ,$$

we can replace this expression by $\overline{W}$ in the formula for Δp. Considering that $q = 1 - p$, we obtain:

$$\Delta p = \frac{pq(\overline{W}_A - \overline{W}_a)}{\overline{W}} .$$

Evolution at the species level: Macroevolution

The sustained effect over generations of all those sources of variation leads to the evolution of the species, which can be displayed

through different outcomes: speciation, extinction, adaptation,[27] co-evolution,[28] etc. Since the main macroevolutionary processes that we are going to consider in this thesis are related to speciation and extinction events, we will focus on these two processes (Freeman and Herron, 2001; Fontdevila and Moya, 2003):

- *Speciation*: It is the process whereby new species[29] arise from a previous one. The main mechanism that leads to a speciation process is reproductive isolation, which avoids the gene flow between two subpopulations from a certain population. The barriers to the gene flow can be of different nature, such as geographical, environmental, ethological, mechanical, or physiological barriers. From the geographical point of view, three basic speciation modes are described (Tamarin, 1996; Fontdevila and Moya, 2003; Gavrilets, 2003):
  - *Allopatric*: This is the classical way of speciation. This speciation process occurs through the appearance of a geographical barrier inside a population that leads to the splitting of the original population into two subpopulations. Over time, this geographical barrier will lead to the divergence of both subpopulations and the origin of two new species.
  - *Parapatric*: In that case, the speciation is given by a geographical isolation but, unlike the allopatric model, in

[27] *Adaptation* is the evolutionary process whereby a population becomes better suited to its habitat. It also denotes the trait that increases the ability of an organism to survive or reproduce, compared to individuals without that trait.

[28] *Coevolution* refers to those correlated evolutionary processes between two interacting species that lead to the reciprocal adaptation of both species through the response of each species to the selection pressure set by the other species.

[29] The meaning of species depends on the biological criterion that is taken into account, as was published by Mayden (1997), who lists two dozen different definitions of species. So, for example, we can find a biological species concept, an ecological species concept, an evolutionary species concept, a morphological species concept, or a phylogenetic species concept, among others.

which the reproductive isolation is sudden, here the reproductive isolation is gradual. This kind of speciation appears, usually, in large distributed populations that contact with a new niche or habitat. There is no physical barrier, but the new habitat constitutes a barrier to the gene flow.

  - *Sympatric*: This model refers to those speciation events that take place in the same range and habitat of the original population. This speciation model is usually related to the origin of a new polymorphism stable for a certain ecological selection, which leads to the divergence of this subpopulation from the original population. This kind of speciation is quite common in parasits.

- *Extinction*: It is a term used to refer to the disappearance of species. Several causes lead to extinction, but in a very general way, we can say that an extinction event has occurred when the last individual of a species die. The extinction events that occur with an uniform rate are called *background extinctions*. Throughout the history of life, several extinction episodes, which were geographically and taxonomically widespread, have been found to be characterized by sudden extraordinary extinction rates (over 60% of the species go extinct), and have been termed *mass extinctions* (Freeman and Herron, 2001). Since the origin of life, six mass extinction events have been identified (Jablonski, 1991; Kareiva, 2004; Wake and Vredenburg, 2008):

  - *Ordovician-Silurian extinction (ca. 439 Mya)*: It led to extinction of 25% of the families and nearly 60% of the genera of marine organisms. The causes were related to big fluctuations in sea level, originated from extensive glaciations, followed by a period of great global warming.
  - *Late Devonian extinction (ca. 364 Mya)*: It eliminated 22% of marine families, and 57% of marine genera. It is con-

sidered to have been related to global cooling after bolide impacts.

- *Permian-Triassic extinction (ca. 251 Mya)*: Aside from the Holocene extinction, this is considered the largest mass extinction event ever. During this extinction, 95% of all species disappeared, including 53% of marine families, 84% of marine genera, and 70% of the land plants, insects and vertebrates. Various causes have been proposed, the most accepted one being the climate change derived from a flood volcanism emanating from the Siberian Traps.
- *End Triassic extinction (ca. 199-214 Mya)*: About 22% of families and 53% of genera of marine organisms were lost. It was related to the opening of the Atlantic Ocean by sea floor spreading associated to massive lava floods that caused significant global warming.
- *Cretaceous-Tertiary extinction (ca. 65 Mya)*: It caused the disappearance of about 16% families, 47% marine genera and 18% vertebrate families. This extinction is responsible for the dinosaur extinction and gave rise to the expansion of mammals and birds. The causes are not clear and different hypothesis have been proposed, two of which are: diverse climate changes derived from volcanic floods in India, and effects derived from a gigant asteroid impact in the Gulf of Mexico.
- *Holocene extinction (Today-ca. 11,000 years ago)*. The increasing human pressure on the environment, since the origin of plant and animal domestication, has derived in the largest extinction event ever. Current extinctions rates are estimated to be 100 to 1000 timer higher than pre-human extinction rates (Pimm et al., 1995). We can exemplify the peril of this situation with the following percentages: 50% of vertebrate animals are classified as threatened, 2.1% of mammals and 1.3% of birds have gone extinct from 1600 to present.

1.1.1 Organism evolution

Reflections about organism evolution were early present in the history of thought (Templado, 1982; Grasa Hernández, 2002). A well documented example is the case of Anaximander (610BC-546BC), who proposed in his work entitled *On Nature* that the first organisms were formed from water and those gave rise to the terrestrial ones (Guthrie, 1962). Like this, different theories about organism evolution were proposed over time, but it was in the 19th century when the theories about organism evolution that greatly influenced contemporary evolutionary biology were proposed. The first relevant evolutionary theory was the *theory of transmutation of species*, proposed by Jean Baptiste Pierre Antoine de Monet, chevalier de Lamarck (1744-1829). This theory postulated that species were created by spontaneous generation but it also states that alteration of some species can cause the appearance of new species (Lamarck, 1809). In 1858, Charles Robert Darwin and Alfred Russel Wallace (Darwin and Wallace, 1858) proposed natural selection as the main driving force of evolution. One year later, Darwin (1859) published his famous *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, where he presents in detail the whole basis of his proposal as theory of evolution.

Although Darwin and Wallace had different ideas of natural selection,[30] both considered evolution by natural selection to be based on four principles (Reznick and Ricklefs, 2009):

- Organisms have individual variations that are faithfully transmitted from parent to offspring.
- All the organisms produce more offsprings than the required to replace themselves in the next generation.

[30]While Darwin emphasized the effect of the competition among individuals of the same species to survive and reproduce, Wallace emphasized the effect of environmental pressure on populations and species, forcing them to become adapted to their local environment.

- Limited resources create a "struggle for existence" that regulates population size, most of the offsprings dying without reproducing.
- The individuals that survive and reproduce are, on average (by virtue of their individual variations), better suited to their local environment than those that do not.

Darwin accepted Lamarck's principle of inheritance of acquired characters as a source of biological variability,[31] and it was only after Darwin's death that the Lamarckian principle of inheritance was denied. Thus, in 1892, August Weissmann provided experimental evidence against soft (lamarckian) inheritance, and postulated his *germ-plasm theory* (Weismann, 1892). This theory states that random mutations are the unique source of change for natural selection to take place. The rejection of the lamarckian inheritance gave rise to an extension of Darwin's theory, coined by George Romanes as *neo-darwinism* (Romanes, 1895). One of the first influential neo-darwinian works was Wallace's *Darwinism*, a defense of natural selection and Weismann's conclusions.

The rediscovery of Mendel's work in the 1890s gave rise to the constitution of genetics as a scientific field, as well as to the works on population genetics. The foundation of population genetics during the 1920s and 1930s led to the proposal of a new theory of evolution that tried to reconcile Darwin's theory with genetics, the *synthetic theory* (Kutschera and Niklas, 2004).

This short sketch of the history of evolutionary biology was considered by Ernst Mayr as a two-phase process (Mayr, 1991): in the first phase, during the 1860s and the 1870s, biologists had to vindicate evolution as a fact, that is, they had to succeed in the explanation that all the organisms were linked in the past through a common set of intermediates. The second phase would have occurred in the

[31] Lamarckian conception of acquired inheritance was labelled by Ernst Mayr *soft inheritance*.

1940s, with the foundation of the modern evolutionary synthesis, when biologists accepted microevolution[32] as a necessary step in evolution. As an example of this confidence in the microevolutionary processes as the basis of evolution, Mayr claimed that (Mayr (1963), pp. 586-587):

> The proponents of the synthetic theory maintain that all evolution is due to the accumulation of small genetic changes, guided by natural selection, and that transspecific evolution is nothing but an extrapolation and a magnification of the events that take place within populations and species [...] essentially the same genetic and selective factors are responsible for evolutionary changes at the species and at the transpecies levels [...] it is misleading to make a distinction between the causes of micro- and macroevolution.

The idea, proposed by Darwin and Wallace, and improved by the synthetic theory, that species evolution is driven through a gradual variance and selection at population level, has led to an intense debate inside evolutionary biology. Using Mayr's historical perspective, we could say that over the last 20-30 years there has been an increasing interest in a third phase of questioning whether microevolutionary processes are enough to explain macroevolution (Penny and Phillips, 2004).

Fossil record provides examples which suggest that morphological evolution was, in general, a gradual process through accumulation of small changes over time. But the fossil record is discontinuous, with a constant presence of certain fossils at each strata, but with

[32]The terms *microevolution* and *macroevolution* were coined by Yuri Filipchenko in 1927 in order to distinguish those evolutionary processes that occur inside a species (microevolution) from those processes that take place among species or higher-level taxa (macroevolution) (Filipchenko, 1927). Those terms were later used by his disciple, Dobzhansky (Dobzhansky, 1937).

temporal transitions of tens of thousands of years between strata. This discontinuity in the fossil record led to the proposal of different evolutionary mechanisms that gave rise to these discontinuities (saltation, punctuated equilibrium, etc). The common premises of those alternative proposals are:

- Evolution is not gradual.
- Microevolutionary processes are not enough to explain macroevolutionary patterns.

The most influential theory in this direction is the *theory of punctuated equilibrium*, proposed by Niles Eldredge and Stephen Jay Gould in 1972 (Eldredge and Gould, 1972). This theory claims that both speciation events and the morphological variations linked to them occur in a short period of time, followed by long periods of *stasis*, i.e. periods of time without apparent change. Although the evolutionary changes in morphology are, perhaps, continuous in the sense of passing through many intermediate stages, they have occurred so rapidly that the fossil record presents the appearance of discontinuous changes. This theory differs from the *saltation hypothesis*, which claims that intermediate stages never existed, the evolutionary discontinuities being due to *macromutations*, i.e. drastic genetic changes that radically alter the phenotype.

1.1.2 Molecular evolution

Until now we have focused on the evolution at organism level but, as we have seen at the beginning of Section 1.1, all those mutations that give rise to variations among individuals are stored at the genome of those individuals. Therefore, how evolution is reflected at the molecular level is the subject of this section.

Molecular evolutionary biology emerged as a scientific field in the mid-1960s, with the amino acid sequencing of hemoglobin, cy-

tochrome c, and other especially abundant proteins in vertebrates. The availability of those datasets allowed for two very influential works in molecular evolution, both based on the comparison of the rate of molecular change among species. On the one hand, Emil Zuckerkandl and Linus Pauling, in 1965, formulated the *molecular clock hypothesis*, based on the observation that the rate of amino acid sequence change for certain proteins appeared to be constant during the diversification of vertebrates (Zuckerkandl and Pauling, 1965). Despite becoming very controversial (Avise, 1994; Hillis et al., 1996), this hypothesis has stimulated much interest in the use of macromolecules in evolutionary studies. Two of the main reasons for this influence are (Li, 1997; Bromham and Penny, 2003; Ho and Larson, 2006; Kumar, 2005):

- If macromolecules evolve at constant rates, they can be used to date evolutionary events.
- The degree of rate variation among lineages can help us to understand the mechanisms behind molecular evolution.

On the other hand, the other influential work was the one published by Motoo Kimura (Kimura, 1968), who, by plotting in time the mutations of the well-studied proteins of human and horses, and extrapolating these evolutionary rates to all of the protein-coding genes in the genome, observed that the mutation rates were far too high to be due to natural selection. This result led him to formulate the *neutral theory of molecular evolution*. By means of this theory, Kimura claimed that most of the mutations that become fixed in populations are neutral, i.e. fixed through genetic drift, while the beneficial mutations fixed by natural selection are extremely rare (Kimura, 1968, 1983).

Both Zuckerkandl's and Kimura's works triggered an intense debate between neutralism and selectionism (Kimura and Ota, 1974; Mayr, 1963; Kreitman, 1996; Ohta, 1996a; Nei, 2005). Most of this debate has

focused on explanations for genetic variation in populations. While neutralists and selectionists agree that deleterious mutations occur frequently in evolving molecules, they profoundly disagree on the relative importance of effectively neutral and beneficial mutations. Neutralists consider that beneficial mutations are rare and are fixed less frequently than neutral or slightly deleterious mutations while, for selectionists, beneficial mutations are abundant (Wagner, 2008a). This controversy between neutralists and selectionists is still not resolved (Ohta, 1992, 1996b; Nei, 2005; Wagner, 2008a; Hurst, 2009), but beyond controversies, neutral theory has become very helpfull as null hypothesis in the detection of natural selection effect on DNA sequences (Li, 1997).

Gene evolution

Development of the sequencing technology in the last decades has provided complete genomes from a large amount of diverse organisms. This availability of genomes has given rise to an increase in the understanding of the evolution of genes and genomes as such. One of the hottest topics in this direction is the comprehension of the formation of new genes (Babushok et al., 2007).

During all these decades, several molecular mechanisms have been described as the basis of gene evolution. Some of such mechanisms are (Mindell and Meyer, 2001; Koonin, 2005; Babushok et al., 2007; Chothia and Gough, 2009):

- *Sequence divergence*: This process basically describes small-scale mutations.
- *Duplication*: Gain of an extra copy of the gene due to large-scale mutation events, like unequal (chromosomal) crossover,[33] se-

[33] *Chromosomal crossover* is one of the final phases of chromosomal recombination, which take place during prophase I of meiosis.

quence duplication, retrotransposition,[34] chromosome duplication or polyploidy (Zhang, 2003; Britten, 2006).

- *Gene fusion*: Combination of pre-existing genes. This can be observed after chromosomal rearrangement phenomena such as unequal crossover, gene conversion,[35] chromosomal transposition, chromosomal translocation or interstitial chromosomal deletion.

- *Horizontal gene transfer*: Process in which an organism incorporates genetic material from another organism, without being the offspring of this organism. It is an important driving force of evolution in bacterias, archaeas, as well as in unicellular eukaryotes (Boto, 2010). The transference of genetic material that takes place during the horizontal gene transfer can be the result of: transference by cell-to-cell (*conjugation*), introduction of foreign genetic material into the cell (*transformation*), or DNA transference via viral infection (*transduction*).

- *Gene loss*: Several mechanisms have been proposed as responsible for the removal of a gene, such as unequal crossover, chromosomal deletion, or chromosomal translocation.

Those genes that are evolutionary related are called *homologs*. In 1970, Walter Fitch coined two of the major forms of homology (Fitch, 1970): *orthologs* (gr. $o\rho\theta o$, 'right'), i.e. those genes diverged through an speciation event, and *paralogs* (gr. $\pi\alpha\rho\alpha-$, 'beside'), i.e. those genes originated from a gene duplication event. Since then, several

[34] *Retrotransposition* is the result of the action of certain reverse transcriptases, *retrotransposons*, which lead to the insertion of intronless copies of genes. One of the best known retrotransposons is LINE-1. The epigenetic effect of retrotransposition has led some biologists to consider retrotransposition as an example of evolutionary mechanism that supports punctuated equilibrium theory (Gogvadze and Buzdin, 2009; Zeh et al., 2009).

[35] *Gene conversion* is an event that occurs during chromosomal recombination. It consists of the transference of DNA sequence from one chromosome to the homologous, the former remaining unchanged.

Homology	Evolutionary process
Orthology	Speciation
Paralogy	Duplication
Xenology	Horizontal gene transfer
Gametology	Barrier to sex chromosome recombination
Ohnology	Whole-genome duplication
Synology	Hybridization of two species

Table 1.2: Main forms of homology (Mindell and Meyer, 2001).

forms of homology have been termed based on the biological process that gives rise to the formation of a new gene (see Table 1.2 (Mindell and Meyer, 2001)).

1.2 Phylogenetic trees: A sketch of evolution

In the previous section we introduced the basic mechanisms by which biological evolution takes place at different organization levels such as genes, populations and species. Throughout the history of thought, a widespread interest in ordering biodiversity has been carried out, with the aim of getting some pattern about how it is organized and so, inferring by which principles it is governed (Templado, 1982; Grasa Hernández, 2002; Kutschera and Niklas, 2004; Ragan, 2009). Since the first evolutionary theory was proposed, the most common way to represent those evolutionary processes has been the tree-like sketches known as *phylogenetic or evolutionary trees*. As an example of this approach, we can consider the case of a population of a certain species. If, inside this population, a genetic barrier appears, impeding the gene flow between both subpopula-

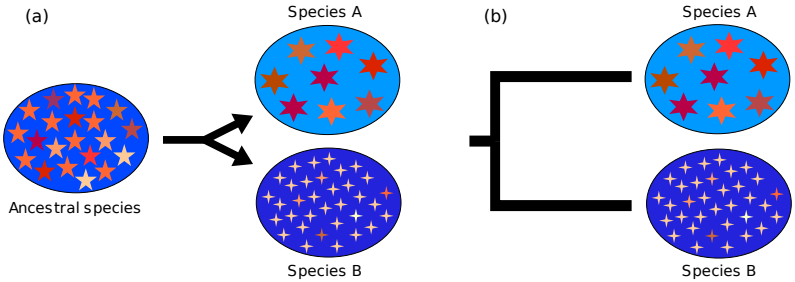


Figure 1.1: Phylogenetic tree as a sketch of evolution. A traditional way to represent the evolutionary history of a group of genes or organisms (a) is with a phylogenetic tree (b).

tions, over time, both ancestral subpopulations will diverge to two different species, species A and B. The way to represent this process would be a phylogenetic tree with three nodes, a *root* and two *tips*. Root and tips represents two different stages in time, where the root corresponds to the ancestral species, and the tips correspond to the species that arose from the speciation event (species A and B) (see Figure 1.1).

In a phylogenetic tree we can distinguish different components (see Figure 1.2) (Li, 1997; Gregory, 2008): root, branches, nodes, tips, etc. The external nodes, referred to as tips or leaves, correspond to existing or extant organisms, which are often called *operational taxonomic units (OTUs)*, a generic term that represents any kind of comparable taxon, such as, for example, individuals or species. In the same way, the term used to refer to the internal nodes, is *hypothetical taxonomic units (HTUs)*, as hypothetical progenitors of the OTUs. A very relevant element in the reconstruction of phylogenetic trees is the *outgroup*, which is not a natural member of the group of interest (*ingroup*), but it represents an OTU identified, by external

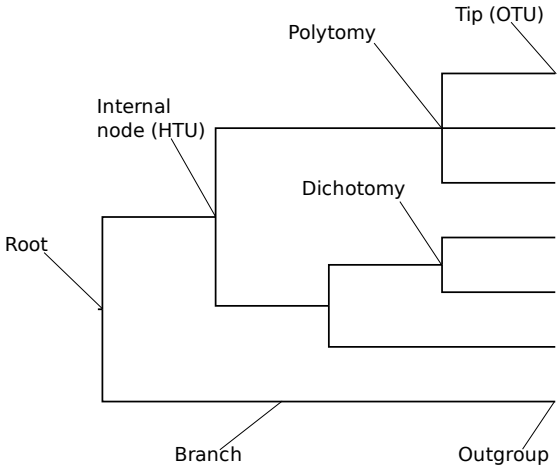
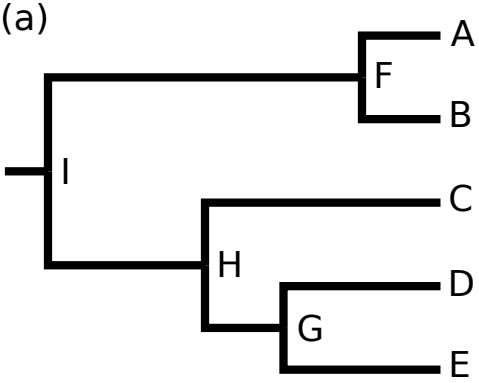


Figure 1.2: Different components of a phylogenetic tree.

information (e.g. paleontological evidence), as branched off earlier than the taxa under study. The outgroup is essential for the rooting of the phylogenetic tree, as well as for the identification of the evolutionary relationship among the ingroup members. Without the outgroup, the tree would remain unrooted. The root is represented as the deepest internal node, and it represents the single common ancestor that the OTUs share.

From the computational point of view, there are different ways of representing phylogenetic trees. So, for example, the classical way to represent the phylogenetic trees in biology is using parentheses and commas, and this format is known as *Newick tree format*, while in complex network theory, the classical way to represent networks is in columns format. In Figure 1.3 we show the representation of a certain phylogenetic tree in both formats, Newick (Figure 1.3(b)) and columns format (Figure 1.3(c)).

(a)



(b)

((A,B)F,(C,(D,E)G)H)I;

(c)

A F
B F
C H
D G
E G
G H
F I
H I

Figure 1.3: Different ways of representing a phylogenetic tree. Representation of a phylogenetic tree (a) using Newick (b) and columns format (c).

1735	C. von Linné: Rank-based classification of living organisms.[36]
1790-1811	W. Smith, G. Cuvier & A. Brogniart: Principle of faunal succession.[37]
1809	J.-B. Lamarck: Theory of transmutation of species, based on increasing complexity and adaptation. Evolutionary tree of animals.[38]
1840	E. Hitchcock: Evolutionary trees, based on paleontology data, of plants and animals, without connection between them.[39]
1859	C. Darwin: A single Tree of Life, with a common ancestor, as a sketch of evolution.[40]
1866	E. Haeckel: Three-kingdom biological classification. First labeled Tree of Life.[41]
1904	G.H.F. Nuttall: Phylogenetic relationships among different groups of animals through conducted precipitin tests of serum protein.[42]
1925	É. Chatton: Two-empire biological classification.[43]
1930s	E. Baldwin: Foundation of comparative biochemistry.[44]
1938	H.F. Copeland: Four-kingdom biological classification.[45]
1944	O. Avery: Identification of the DNA as the genetic material.[46]
1950	W. Hennig: Foundation of phylogenetic systematics.[47]
1955	F. Sanger: Complete sequencing of insulin.[48]
1958	R.R. Sokal & C.D. Michener: UPGMA method.[49]
1963	E. Margoliash: Cytochrome c phylogeny for horse and other species.[50]
1962	E. Zuckerkandl & L. Pauling: Molecular clock hypothesis.[51]
1966	R.V. Eck & M.O. Dayhoff: Maximum parsimony method.[52]
1967	L.L. Cavalli-Sforza & A.W.F. Edwards: Maximum likelihood method.[53]
1969	R.H. Whittaker: Five-kingdom biological classification.[54]
1976	W. Fiers et al.: First whole-genome (bacteriophage MS2) sequenced.[55]
1977	C. Woese: Six-kingdom biological classification system.[56]
1983	K. Mullis: Invention of the PCR.[57]
1986	J. Gauthier: First published work based on phylogenetic nomenclature.[58]
1987	N. Saitou & M. Nei: Neighbor-Joining method.[59]
1990	C. Woese: Three-domain biological classification.[60]
1996	B. Rannala & Z. Yang, B. Mau et al. & S. Li: Bayesian inference of phylogeny.[61]
2000	P.D. Cantino & K. de Queiroz: First public draft of PhyloCode.[62]
2004	T. Cavalier-Smith: Six-kingdom biological classification.[63]

Table 1.3: Some of the main evens in the history of phylogenetics.

1.2.1 Kinds of evolutionary trees

There are different kinds of evolutionary trees, depending on the sort of evolutionary event to be represented. Thus, they are (Gregory, 2008; Avise, 2009):

- *Cladogram*: Evolutionary tree that represets the evolutionary relationships only, without taking into account evolutionary distances.

[36]See von Linné (1758).
[37]See Winchester (2001); Cuvier and Brongniart (1822).
[38]See Lamarck (1809).
[39]See Hitchcock (1840).
[40]See Darwin (1859).
[41]See Haeckel (1866).
[42]See Nuttall (1904).
[43]See Chatton (1925).
[44]See Baldwin (1937).
[45]See Copeland (1938).
[46]See Avery et al. (1944).
[47]See Hennig (1950).
[48]See Ryle et al. (1955).
[49]See Sokal and Michener (1958).
[50]See Margoliash (1963).
[51]See Zuckerkandl and Pauling (1962).
[52]See Eck and Dayhoff (1966).
[53]See Cavalli-Sforza and Edwards (1967).
[54]See Whittaker (1969).
[55]See Ryle et al. (1955).
[56]See Balch et al. (1977); Woese and Fox (1977).
[57]See Mullis (1990).
[58]See Gauthier (1986).
[59]See Saitou and Nei (1987).
[60]See Woese et al. (1990).
[61]See Rannala and Yang (1996); Mau (1996); Li (1996).
[62]See Cantino and de Queiroz (2000).
[63]See Cavalier-Smith (2004).

	Stepwise clustering	Exhaustive search
Distance Matrix	UPGMA	Fitch-Margoliash
	Neighbor-joining	
Character State		Maximum parsimony
		Maximum likelihood
		Bayesian inference

Table 1.4: Most commonly used reconstruction methods (adapted from Salemi and Vandamme (2003)).

- *Phylogram*: Evolutionary tree that represets the evolutionary relationships, taking into account evolutionary distances based on some character (genetic distance, morphological distance, etc).

- *Chronogram*: Evolutionary tree that represets the evolutionary relationships, including evolutionary distances based on time (e.g. millions of years).

1.2.2 Phylogenetic tree reconstruction methods

Since the publication, in 1958, of the *unweighted pair-group method with arithmetic mean*, known as *UPGMA* (Sokal and Michener, 1958), a large amount of methods for the reconstruction of phylogenetic trees have been proposed. The different methods can be grouped according to two basic criteria (see table 1.4) (Salemi and Vandamme, 2003; Lemey et al., 2009): (1) Whether they use distance matrix of pairwise dissimilarities (*distance matrix methods*) or they use discrete character states (*character-state methods*); and (2) whether they cluster OTUs stepwise, inferring only one best tree (*stepwise clustering methods*), or they consider all theoretically possible trees (*exhaustive search methods*).

On the one hand, distance matrix methods define the phylogenetic relationships based on the pairwise distance matrix obtained from the measure of dissimilarities of each pair of OTUs. Those methods are specially appropriate for analyzing sequence data, the evolutionary distances being usually measured in numbers of nucleotides or amino acid substitutions between sequences. These evolutionary distances are calculated using evolutionary models that allow for the correction of the percentage of difference between sequences. Since the distance methods discard the original character state of the taxon, the reconstruction of the character states of the ancestral nodes is not possible. The main advantage of these methods is that they are much less computer-intensive. On the other hand, character-state methods can be used with any set of discrete characters, such as morphological characters, physiological properties, restriction maps, or sequence data, and each character is analyzed separately and usually independently from the other characters. In the case of sequence use, the character is defined as each position of the aligned sequence. Since those methods retain the original character status of the taxon, character-state methods are useful in the reconstruction of the character state of the ancestral nodes.

Stepwise clustering methods infer only one best tree starting the tree reconstruction by examining the local subtrees. Therefore, the most closely related OTUs are combined to form a cluster, and this cluster is treated as a single OTU, representing the ancestor of the OTUs it replaces. And this process is repeated for the next closest OTUs and so on. The way to determine the relationship between OTUs differs from one stepwise clustering method to the other. These methods are usually fast and are able to accommodate large numbers of OTUs. Since they infer only one best tree, the confidence in the correctness of an inferred tree has to be estimated through supplementary statistical methods. Otherwise, the phylogenetic tree reconstruction by exhaustive search methods considers all the theoretically possible trees and selects the best one by certain criteria. The main drawback of these methods is that the computing time grows fast with the

number of taxa, being the number of bifurcated rooted trees for n OTUs: $\frac{(2n-3)!}{(2n-2(n-2))!}$. This means that for a dataset larger that 10 OTUs (34,459,425 possible rooted trees), only a subset of possibles trees can be examined. Hence, several strategies are used in order to search the so-called *tree space*, but there is no algorithm that guarantees that the best possible tree was actually considered.

Most of the distance matrix methods use stepwise clustering, while most of the character state methods use exhaustive search approach. The main distance matrix methods are UPGMA, neighbor-joining, Fitch-Margoliash, while the main character state methods are maximum parsimony, maximum likelihood, bayesian inference (Li, 1997; Page, 1998; Lemey et al., 2009).

- *UPGMA (unweighted pair group method with arithmetic mean)*. It is the first and simplest method for phylogenetic tree reconstruction by distance matrix data (Sokal and Michener, 1958). It was originally proposed for phenotypic distance matrix data, but nowadays it is also used for sequence-based phylogenetic tree reconstruction. UPGMA assumes a constant rate of evolution, hence it tends to give the wrong tree when evolutionary rates are not constant. For the tree reconstruction, it uses a stepwise clustering algorithm by which the phylogenetic relationships are inferred in order of decreasing similarity. In that sense, those OTUs with closest similarity are the first identified, and so on. After each local clustering, the distances between the new cluster and the remaining OTUs are redefined, the distance of the newly formed cluster corresponding to the average of the distances of the original OTUs.

- *Fitch-Margoliash (FM)*. It is an exhaustive search distance matrix method. It uses a weighted least square algorithm, based on genetic distance, for the evaluation of all the possible trees for the shortest overall branch length (Fitch and Margoliash, 1967).

- *Neighbor-joining (NJ)*. It is a stepwise clustering method that, like UPGMA, uses distance matrix data. This algorithm defines the phylogenetic relationships by minimizing the total length of the tree (Saitou and Nei, 1987). The method starts with a star-like tree without internal branches. The first step consists of separating the first pair of OTUs from the remaining of OTUs, and measuring the length of the resulting tree. The algorithm repeats this process for each OTU till the shortest tree is obtained.

- *Maximum parsimony (MP)*. It is an exhaustive search method whose main principle is to reconstruct the tree that requires the smallest number of character changes. The approach was first developed for amino acid sequence data (Eck and Dayhoff, 1966), and it was later that the method was modified for nucleotide data (Fitch, 1977). The algorithm infers all the possible tree topologies and infers, for each topology, the minimum number of character changes needed to explain all the nodes of the tree. Since more that one tree can have the minimum number of nodes, the algorithm does not necessarily infer a unique tree topology.

- *Maximum likelihood (ML)*. Like MP, ML is an exhaustive method that uses discrete character data, but in this case, the best tree is the most likely, based on an evolutionary model. The first application of the approach was developed for tree reconstruction through gene frequency data (Cavalli-Sforza and Edwards, 1967), and was later applied to amino acid (Felsenstein, 1973) and nucleotide sequence data (Felsenstein, 1981). The algorithm calculates the likelihood for each tree, based on the probability of observing that tree given a certain evolutionary model. After obtaining the likelihood of all the tree topologies, the most likely tree is chosen as the best one. ML is able to capture all the information that the data tell us about the phylogeny under a certain model but, as a drawback, the algorithm is computationally very demanding.

- *Bayesian inference (BI)*. This approach is closely related to the ML approach. But, while the ML algorithm maximizes the probability of observing a certain tree, the BI approach maximizes the posterior probability. For a certain evolutionary model, the posterior probability of a tree is proportional to the likelihood of that tree, multiplied by the prior probability, which is the probability of the model without any reference about the data (Li, 1996; Mau, 1996; Rannala and Yang, 1996).

1.2.3 Challenges of the evolutionary trees: anagenesis, polytomies and reticulate evolution

The example that we used at the beginning of this section for the explanation of the basic fundamentals of phylogenetic trees represents a standard example, but it does not imply that all the evolutionary processes that take place can be represented in the same way. Some examples of those non-standard cases are the following:

- *Anagenesis*. When we explained the speciation process, we focused on cladogenetic processes, where divergence between subpopulations inside a certain population gives rise to two or more daughter species. But in some cases, evolution takes place homogeneously in all the members of the species so that, after a certain time, if we compare the actual species with the ancestral one, we could not classify the actual species as the same species as the ancestral one. So, rather than giving rise to two or more species, the speciation process would give rise to a single new species. This speciation event is called *anagenesis* (Tamarin, 1996). It implies a change in the way to represent these speciation events in a phylogenetic tree. In fact, a distinction is made in cladistics between cladogram and evolutionary tree. In cladograms taxa are always represented as tips of the tree, without taking into account if the taxa are extant or extinct, or whether one or more of the taxa are an-

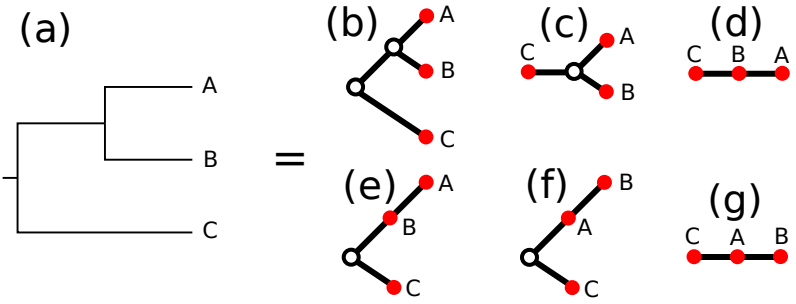


Figure 1.4: Anagenesis in evolutionary trees. For a given cladogram (a) there are six different evolutionary trees consistent with the cladogram (b-g), considering the anagenesis events (c-g) (Page, 1998).

cestral to any of the others. However, in an evolutionary tree some of the taxa may be ancestral to the others. Therefore, an event of anagenesis, rather than being represented by a bifurcation in the phylogenetic tree, would be represented as a chain, where the original species would be represented as an internal node, instead of being represented as an external node (see Figure 1.4).

- *Polytomies*. The example case that we depicted in Figure 1.1 was a binary tree, but in some cases, the branching events are not necessarily binary but polytomic. Polytomies can be (Maddison, 1989; Purvis and Garland, 1993):

  - *Soft polytomies*. The main goal of the reconstruction methods is to infer a fully resolved phylogeny, but a common problem in the reconstruction of phylogenies is the presence of artifacts derived from the inference, due to contradictory results from conflicting data and lack of in-

formation about the real evolutionary relationship. This kind of artifacts leads to the presence of *soft polytomies*.

  - *Hard polytomies*. Those polytomies are due to diversification events that take place simultaneously, e.g. adaptive radiations.

- *Reticulated evolutionary events*. As we have seen all over this section, evolutionary processes have been traditionally represented as tree-like processes, but reticulated evolutionary events are quite common in nature, like hybridization or lateral gene transfer between species, or tokogenetic relationships in a population with sexual reproduction. These reticulated events require a different perspective to represent the evolutionary processes, as well as a consequent proposal of alternative methods for the reconstruction of *phylogenetic networks* (Posada and Crandall, 2001; Morrison, 2005; McBreen and Lockhart, 2006).

1.2.4 Organism phylogenies

As we commented on in Section 1.1.1, the 19th century implied a turning point in the history of evolutionary thought with Lamarck's (Lamarck, 1809) and Darwin/Wallace's (Darwin and Wallace, 1858; Darwin, 1859) theories of biological evolution. Derived from this intellectual context, in order to reflect the transmutation processes that take place during evolution, several tree-like diagrams were proposed, the most influential ones being those proposed by Lamarck (1809), Hitchcock (1840), Darwin (1859) and Haeckel (1866) (for a comprehensive historical review see Ragan (2009)). This tree-like representation of the whole history of biological evolution was termed *Tree of Life* (see Figure 1.5).

The Tree of Life tries to reflect the evolution of all the Earth living forms (Cracraft and Donoghue, 2004). The main core of the Tree of Life represents the macroevolutionary processes, that is, those

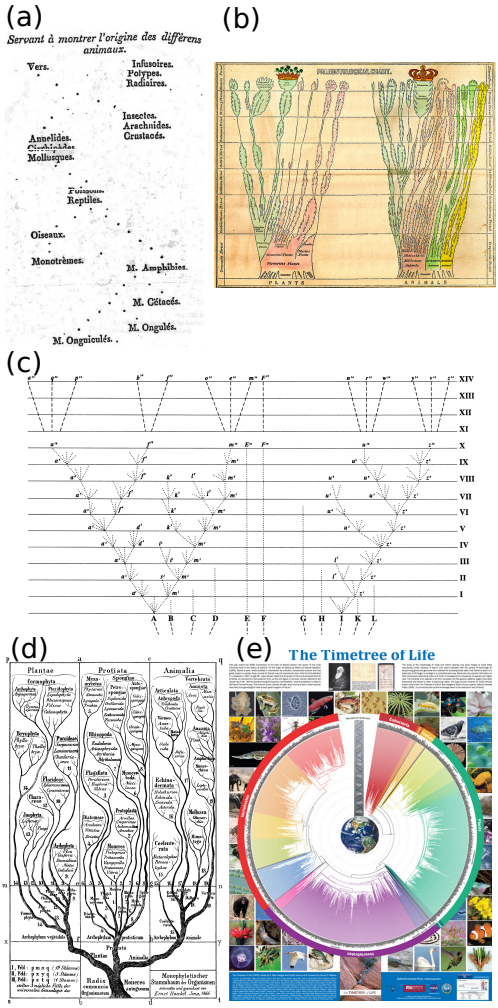


Figure 1.5: Historical evolution of the illustration of the Tree of Life. Lamarck (1809) (a), Hitchcock (1840) (b), Darwin (1859) (c), Haeckel (1866) (d) and Timetree (2010) (e).

processes that take place among species or higher-level taxa, but the surface of the tree, reflects all the microevolutionary events that take place among populations inside each species (Herrada et al., 2008).

This tree-like way to consider the evolution of life has given rise to different controversies. Some of the most relevant ones are:

- *Is the Tree of Life really tree-like?* Although Darwin's works reinforced the idea of sketching evolution as a Tree of Life, this tree-like metaphor is very discussed. In fact, before the Tree of Life metaphor, networks were proposed for depicting evolution, and since the proposal of the first evolutionary theories, both approaches have coexisted as different ways of conceiving evolution (Ragan, 2009). Thus, during the last decades the number of publications describing reticulatory evolutionary processes and proposing alternative models of representing the evolutionary history of all organisms on Earth has increased. In this line, for example, the Net of Life (Ragan et al., 2009) and the Ring of Life (Rivera and Lake, 2004; Rivera, 2007) constitute alternative metaphors to the Tree of Life. The former tries to highlight the presence of reticulatory events over the history of life, like horizontal gene transfer, hybridization, etc, while the latter makes the emphasis on the reticulatory events between Eubacteria and Archaea that could have given rise to the origin of eukaryotes.

- *Is the Tree of Life a single-rooted tree?* One of the most highlighted ideas from Darwin's *On the origin of the Species* was the belief on a last common ancestor from which all the living forms would have evolved. However, studies aimed at determining the evolutionary relationship among Eubacteria, Eukarya and Archaea have led to question this single-rooted Tree of Life proposed by Darwin, in favor of a multiple-rooted Tree of Life where. Instead of a last universal common ancestor, a common ancestral community of primitive cells would have given rise

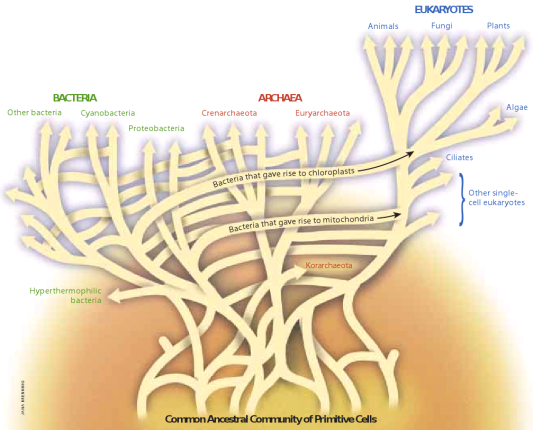


Figure 1.6: Tree of Life rooted at a common ancestral community of primitive cells, instead of a single last universal common ancestor (Doolittle, 2000).

to Earth living forms (Doolittle, 2000; Steel and Penny, 2010) (see Figure 1.6).

- *Nomenclature criteria*. Until the 20th century, the main way of finding some logical order in biodiversity was supported by the classification of the organisms based on their similarity by resemblance. This was the main principle of biological taxonomy, whose main precursors were Aristotle (384-322 BC) (Aristóteles, 1994) and Carl von Linné (1707-1778). von Linné (1758) proposed a classification based on six main hierarchical categories or ranks (from an upper to a lower level: *kingdoms*, *classes*, *orders*, *genera*, *species*, *variety*).[64] Years later,

[64]Traditionally, eight main taxonomic ranks are defined: *domain* (e.g. Eukarya), *kingdom* (e.g. Animalia), *phylum* (*division* in Botany) (e.g. Chordata), *class* (e.g. Mammalia), *order* (e.g. Carnivora), *family* (e.g. Canidae), *genus* (e.g. *Canis*), *species* (e.g. *Canis lupus*) (de Queiroz, 1997). Nowadays, taxonomic nomenclature is regulated

> Darwin's evolutionary revolution entailed a change in the way of classifying organisms, so that instead of reflecting the degree of similarity by resemblance, the hierarchical classification should be based on the degree of similarity by descendant, reflecting the evolutionary relationships among organisms. So, Hennig's *Grundzüge einer Theorie der phylogenetischen Systematik* (1950) gave rise to the foundation of *phylogenetic systematics*, the biological classification methodology based exclusively on evolutionary relationships (Hennig, 1950, 1966). Nowadays, most of the biologists agree that biological classification should be based on the phylogenetic relationships of the organisms (Dubois, 2007). Nevertheless, ever since the change of paradigm, with the inclusion of evolutionary thought in biological taxonomy, there has been an intense debate among those biologists who give preeminence to the pre-evolutionary systems of biological classification, and those who expect classifications to reflect modern evolutionary and phylogenetic findings. The main core of this discussion is based on the fact that the classification proposed by the former claims the use of categorical ranks, taxonomic categories, while the latter claim that the classification should be entirely evolution-based, thus rank-free (de Queiroz, 1988, 1997; Benton, 2000; Nixon and Carpenter, 2000; Bryant and Cantino, 2002; Keller et al., 2003; de Queiroz, 2005; Rieppel, 2005, 2006a,b; Hillis, 2007; Ereshefsky, 2007).

With the development and improvement of the different reconstruction methods, a myriad of datasets of phylogenetic trees have been published. In order to make this large amount of data user friendly,

---

by the specific *Nomenclature Codes* (International Code of Zoological Nomenclature (ICZN) (Ride et al., 1999), International Code of Botanical Nomenclature (ICBN) (McNeill et al., 2007), International Code of Nomenclature of Bacteria (ICNB) (Lapage et al., 1992), International Committee on Taxonomy of Viruses (ICTV) (Fauquet et al., 2005)), which allow classifications divided into an indefinite number of ranks.

different consortia and databases have been created. Some of the most important ones are:

- *TreeBASE*. It is the main phylogenetic tree database. TreeBASE is fed from user-submitted phylogenetic trees and includes trees of species, populations and genes (TreeBASE, 2010; Sanderson et al., 1994). In June 2010, TreeBASE contained 6,500 trees.

- *Tree of Life Web Project*. It is a collection of information for every species and for each group of organisms, living or extinct. It is built with the collaboration of hundred of experts and amateur contributors. The structure of the web page follows the phylogenetic branching pattern between groups of organisms (Tree_of_Life_Web_Project, 2010; Maddison et al., 2007).

- *TimeTree*. This database provides information about the timescales of the evolutionary processes all over the Tree of Life (Timetree, 2010; Hedges et al., 2006; Hedges and Kumar, 2009).

- *Catalogue of Life*. It is a taxonomic collection with the classification of the organisms on Earth (Catalogue_of_Life, 2010; Bisby et al., 2010). The last version, Catalogue_of_Life (2010), contains 1,257,735 species.

Apart from these, there are several databases focused on the partial reconstruction of the Tree of Life focusing on a certain group of organisms, such as: The_Green_Tree_of_Life (2010), FLYTREE (2010), Assembling_the_Fungal_Tree_of_Life (2010), AmphibiaTree (2010), The_Beetle_Tree_of_Life_Project (2010), HymAToL (2010), NemATOL (2010), Early_Bird (2010), Cypriniformes_Tree_of_Life (2010), The_Mammal_Tree_of_Life (2010), Phylogeny_of_Spiders (2010), etc.

1.2.5 Gene phylogenies

As we commented on in Section 1.1.2, the genes that share a common ancestor are considered homolog genes and constitute a *gene family* [65] (Dayhoff, 1965-1978). The representation of a gene family through a phylogenetic tree is labeled *gene phylogeny*.[66] Unlike in the case of organisms, where the idea of a global Tree of Life that integrates all the Earth living forms is widely accepted, the fact that the different gene families do not share any last universal common ancestor makes the integration of the different gene families in a global "Tree of Life of genes" unrealistic.

In the last decade we have witnessed a great progress in the characterization of the gene families. A sign of all this progress is the large amount of databases of gene phylogenies that has been created in this period. Some of the more relevant ones are:

- *PANDIT*. It is a collection of protein phylogenies based on the protein families database *Pfam*(Pfam, 2010). PANDIT includes the phylogenies of almost 8,000 protein families (PANDIT, 2010; Whelan et al., 2003, 2006).[67]

- *TreeFam*. It is a database focused on animal gene phylogenies, with more that 16,000 gene families (TreeFam, 2010; Li et al., 2006; Ruan et al., 2008).

[65] Those gene families of genes that encode for proteins are usually referred as *protein families* (Dayhoff, 1965-1978). This term is often used as a nearly synonym of gene family.

[66] A naïve expectation of molecular systematics is that gene phylogenies match organism phylogenies, i.e. obtaining the first would necessarily give us the second (Page, 1998). However, as we have seen in Section 1.1.2, speciation is not the only evolutionary mechanism that takes part in gene evolution. So, the higher the presence of the other kinds of homologies (paralogy, xenology, etc) in the evolution of a certain gene family, the more discrepancies between gene phylogenies and organism phylogenies will be found.

[67] Due to problems with the funding support, PANDIT is frozen since November 13th 2008.

- *PhylomeDB*. This database provides all the gene phylogenies that integrate a genome (phylomes) for human (157,233 trees), *E. coli* (9,280 trees) and *S. cerevisiae* (5,811 trees) (PhylomeDB, 2010; Huerta-Cepas et al., 2008).

In addition to these databases, there are some others like: (GreenPhylDB, 2010; SYSTERS, 2010; HOVERGEN, 2010), etc.

1.3 Complex network theory and evolutionary biology

The availability of a large amount of data, as a result of the rising of high-throughput techniques in biology, has entailed the search for new statistical approaches that allow the analysis and interpretation of those datasets (Levchenko, 2001; Proulx et al., 2005; Kwoh and Ng, 2007; Almaas, 2007). Such is the case of the application of complex network theory to the screening of data obtained from biological studies. So, complex network theory has been applied to diverse fields from biology (Boccaletti et al., 2010) such as the characterization of molecular networks (Kwoh and Ng, 2007; Almaas, 2007), the analysis of ecological networks (Ings et al., 2009), or the study of neural networks (Bullmore and Sporns, 2009). In the following section I will give a short introduction to complex network theory, as well as a review of the most important applications of the complex network theory to evolutionary biology.

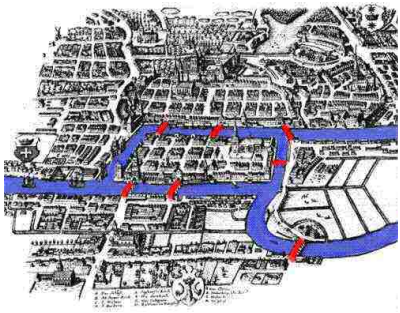


Figure 1.7: Map of Königsberg (current Kaliningrad, Russia) in Euler's time showing the location of the seven bridges that inspired the *Königsberg bridge problem*. The upper bridges (from left to right): Krämer, Schmied, Holz. The central bridge: Honig. The lower bridges (from left to right): Grün, Köttel, Höhe.

1.3.1 Complex networks: The skeleton of complex systems

In 1736 the mathematician Leonhard Euler presented to the St. Petersburg Academy the solution to the *Königsberg bridge problem*,[68] published in 1741 (Euler, 1741) (see Figure 1.7). The resolution of this problem entailed the foundations of graph theory, the branch of discrete mathematics that would be, till the foundation of complex network theory in 1998, the main responsible for the study of networks. More than two centuries after Euler's work, Paul Erdös and Álfred Rényi published a very influential work on random graphs (graphs in which each pair of nodes is connected with a probability

[68] Königsberg bridge problem is a notable historical problem in mathematics, consisting of finding a round trip that traverse each of the bridges of the Prussian city Königsberg (now Kaliningrad, Russia) once and only once. Euler proved that it does not exist any route able to cross each of the seven bridges only once.

1736	L. Euler: Solution to the Königsberg problem.
1929	F. Karinthy: Small-world hypothesis.
1959	Erdös & Reny: First work on random graphs.
1967	S. Milgram: Small-world experiment.
1998	D. Watt & Strogatz: Small-world networks.
1999	A.-L. Barabási & R. Albert: Scale-free networks.

Table 1.5: Some of the main evens that led to complex networks theory foundation.

p) (Erdös and Rényi, 1959). This work meant the introduction of probabilistic methods in graph theory, and the foundation of random graph theory. But it was late in the 1990s when the studies on networks underwent an extraordinary impulse, with the publication on small-word networks by Duncan J. Watts and Steven Strogatz in 1998 (Watts and Strogatz, 1998) and the paper by Albert-László Barabási an Réka Albert on scale-free networks one year later (Barabási and Albert, 1999). Both works involved the rise of complex network theory as a scientific field.

Based on the fact that a complex network constitutes the skeleton of a complex system, complex network theory arises from the application of the classical graph theory to the comprehension of complex systems (Albert and Barabási, 2002). The simplest way to characterize a complex system is through the concept of *emergence*. A complex system is a system whose behavior cannot be defined through the individualized description of each component. In that sense, a complex system can be defined by means of the classical sentence: "The whole is more that the sum of their parts" (Anderson, 1972). That means that the way in which the different components of the complex system interact must be taken into account in order to understand its global behavior. This kind of behavior gives way to *emergent phenomena*, which constitute the essence of a complex sys-

tem. This conception of what a complex system is leads us to an easy understanding of why a complex network represents the skeleton of a complex system. As a classical graph, a complex network is defined by nodes that are interacting through links. The fact that a complex network is a representation of the interactions that take place in a complex system makes a complex network crucial in the understanding of a complex system, in the sense that it helps us in the analysis of the key elements of the complex system that derives in emergent phenomena, i.e. the interactions among the different parts (Mitchell, 2009).

1.3.2 Complex networks in evolutionary biology

Since its foundation, complex network approach has been extensively applied to the study of biological evolution (Lässig and Valleriani, 2002; Képès, 2007; Junker and Schreiber, 2008; Boccaletti et al., 2010). From the different levels of organization of biological systems in which biological evolution is reflected,[69] the ones in which application of complex networks theory has been especially successful are (see Figure 1.8):

- *Cellular level*. As the nodes of networks at this level are representing molecules (genes, proteins, metabolites, etc), these networks can be also referred to as *molecular networks*. Different sorts of networks can be defined depending on the nature of nodes and links. Thus, we can find protein-protein interaction networks, metabolic networks and gene-regulatory networks, among others. The complex network approach has made it possible to obtain a global picture of the biological processes that happen inside a cell. The availability of this

[69] The standard levels of biological organization are (from the lower level to the highest one): molecules, organelles, cells, tissues, organs, organ systems, organisms, populations, communities, ecosystems and biosphere (Brown, 1995).

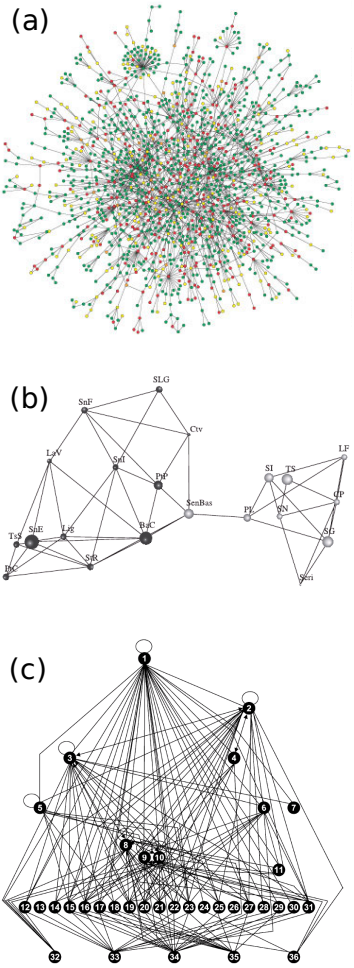


Figure 1.8: Some examples of the application of complex network approach to biological systems: (a) molecular network (Jeong et al., 2001), (b) population network (Dyer and Nason, 2004), (c) ecological network (Woodward et al., 2005b).

global framework has enabled the addressing of relevant topics from the evolutionary point of view (Proulx et al., 2005; Stumpf et al., 2007), such as molecular network evolutionary dynamics (Barabási and Albert, 1999; Barabási et al., 1999; Solé et al., 2002; Wagner, 2003; Berg et al., 2004) (for a comprehensive review, see (Yamada and Bork, 2009)), network robustness (Jeong et al., 2001; Ghim et al., 2005; Kim et al., 2007; Suthers et al., 2009), or correlation between gene relevance and mutation rate (Fraser et al., 2002; Jordan et al., 2003; Hahn et al., 2004; Fraser, 2005; Hahn and Kern, 2005; Zhou et al., 2008).

- *Population and metapopulation level*. If we consider a population based on their individuals and the genetic relationships among them, we obtain a *population network* (Dyer and Nason, 2004). Population networks have been extensively used in population genetics and in phylogeographical studies (Garrick et al., 2010). Last decade has witnessed an increasingly presence of the complex network approach in population genetic studies (Dyer and Nason, 2004; Garrick et al., 2010). Such is the case of the works published by Dyer and Nason (2004); Dyer (2007); Giordano et al. (2007); Rozenfeld et al. (2007, 2008); Garrick et al. (2009).

- *Ecosystem level*. The networks that arise from the ecological interactions among organisms inside an ecosystem are known as *ecological networks*. Three main types of ecological networks are described (Ings et al., 2009): *"traditional" food webs* (based on the trophic interactions among organisms) (Dunne et al., 2002; Montoya and Solé, 2002; Brose et al., 2005; Woodward et al., 2005a,b; Brose et al., 2006; Montoya et al., 2009; Woodward, 2008), *host-parasitoid networks* (focused on the special trophic interactions between parasitoids and their hosts) (Müller et al., 1999; Lewis et al., 2002; Morris et al., 2004; Vázquez et al., 2005; Bukovinszky et al., 2008; Veen et al., 2008) and *mutualistic networks* (those which are specialized on ecosystem services such as pollination and seed dispersal, rather than on population

dynamics or energy fluxes *per se*) (Jordano et al., 2003; Blüthgen et al., 2004, 2006, 2007; Basilio et al., 2006; Montoya et al., 2006; Waser, 2006; Vázquez et al., 2009). Together with the works on molecular networks, these are the most important fields in biology where the application of complex network theory has proved to be very successful.

1.3.3 Complex tree-like networks in evolutionary biology

In the previous section, we have focused on applications of complex network theory to network-like biological systems but, as we have commented on in Section 1.2, there are evolutionary processes that can be represented in a tree-like network.

Tree-like networks are widely present in nature. For example, we can find them in vascular plants, vascular tissues, river basins, etc. (Ball, 2009) (Figure 1.9). The first steps in the understanding of branching patterns were carried out by Leonardo da Vinci (1452-1519), who suspected that there should be some rules governing tree growth (Figure 1.10). Da Vinci's work on branching processes was improved at the end of 19th century by Wilhelm Roux, a pupil of Ernst Haeckel, based on his works on blood vessels (Roux, 1878). Later, in the 1920s, Cecil Murray applied Roux's rules to plant branching, and proposed a mechanism that was able to explain the branching growth of vascular plants (Murray, 1926). The branching mechanism that he proposed was based on a parsimonious minimization principle. Half a century later, in the 1970s, Luna Leopold, inspired by the analogies in form and function between rivers and biological branching networks as fluid-distributing systems, extrapolated Murray's principle to river basins (Leopold, 1971).

Within a complex network context, the analysis of branching processes have been extensively carried out till became an outstanding field (West et al., 1997; Banavar et al., 1999; West et al., 1999; Banavar

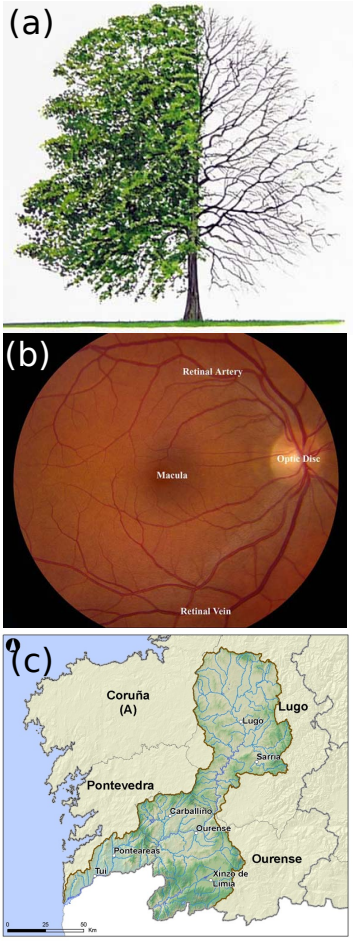


Figure 1.9: Some examples of tree-like networks: (a) Quercus pirenaica (cubiFOR, 2010), (b) retinal blood vessels (Webvision, 2010), (c) Rio Miño basin (Confederacion_Hidrografica_del_Mino-Sil, 2010).

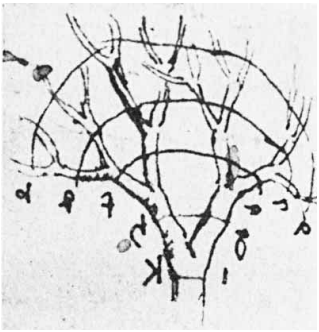


Figure 1.10: Leonardo da Vinci's sketch representing the branching pattern of trees. He depicted that the total thickness of branches along each of the arcs would equal the thickness of the trunk (Richter, 1939).

et al., 2002; West and Brown, 2005; Makarieva et al., 2005; Banavar et al., 2006). The popularity of the complex network approach is such that it has been extended to the characterization of very diverse systems such as ecological tree-like networks (Garlaschelli et al., 2003; Camacho and Arenas, 2005; Zhang, 2009; Zhang and Guo, 2010) or phylogenetic trees (Campos and de Oliveira, 2004; Herrada et al., 2008).

1.3.4 Basic concepts in network theory useful to analyze trees

A *tree-like complex network* is a specific instance of complex network which is characterized by the following properties (Celko, 2004):

- Absence of loops or closed paths.

- Every two nodes in the tree are connected by one (and only one) path.
- The number of links is one less than nodes it has.

As a complex network, the mathematical definition of a tree-like network is a pair of sets, $G = \{V, E\}$, where V is a set of nodes (or vertices), and E is a set of links (or edges) in which each link connects a couple of nodes. Tree-like networks can be *directed* or *undirected*. In directed networks, the interaction from node i to node j does not necessarily comprise an interaction from j to i. On the contrary, when the interactions are symmetrical, we say that the network is undirected. Moreover, networks can also be *weighted* (Almaas et al., 2005). A weight is defined as a scalar that represents the strength of the interaction between two nodes. In an unweighted network, instead, all the edges have the same weight (generally set to 1) (Castelló, 2010).

During the last decade, a large amount of measures have been proposed in order to characterize the topological properties of the complex networks. Below we describe some of those measures which are useful for the characterization of tree-like networks (Albert and Barabási, 2002; Newman, 2003; Almaas et al., 2005; da F. Costa et al., 2007; Zhang et al., 2007):

- *Weight distribution*. For weighted networks, the *weight distribution*, $P(w)$, measures the probability that a randomly selected edge has exactly weight w.
- *Degree distribution*. The *degree* of a node i, k_i, is the number of links connected to that node, and the fraction of nodes in a network with degree k is expressed by the *degree distribution*.

  A natural generalization of the degree of a node in weighted networks corresponds to the *node strength*, s_i, which represents the sum of weights for all the links j connected to a node i, w_{ij}.

The *strength distribution*, $P(s)$, expresses the fraction of nodes with strength s.

- *Average degree*. The *average degree* of a network corresponds to the average of k_i for all nodes in the network.

- *Degree-degree correlation*. Correlations between the degrees of different vertices has been found to play an important role in many structural and dynamic network properties. The most natural approach is to consider the correlation between two nodes connected by a link. This correlation can be expressed by the *joint degree distribution*, $P(k,k')$, the probability that an arbitrary link connects a node of degree k with a node with degree k'. In terms of conditional probability, this could be expressed as the probability that an arbitrary neighbor of a node of degree k has degree k'. This conditional probability can be computed in the following way (Boguñá and Pastor-Satorras, 2002):

$$P(k'|k) = \frac{\langle k \rangle P(k,k')}{kP(k)} .$$

An interesting quantity related to degree correlations is the *average degree of the nearest neighbors for nodes with degree k*, $k_{nn}(k)$, which is given by:

$$k_{nn}(k) = \sum_{k'} k' P(k'|k) .$$

- *Average path length*. The distance between two nodes i and j, d_{ij}, is given by the *shortest path length*, i.e. the number of links along the shortest path connecting them. An important quantity that depends on the overall network structure is the *average path length*, which is defined as the mean value of d_{ij}. Thus, for a directed network of N nodes, the average path length is:

$$l = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \, .$$

A closely related measure to the average path length is the so-called *global efficiency*, E (Latora and Marchiori, 2001). Assuming that the efficiency for sending information between nodes i and j is proportional to the reciprocal of their distance, this measure quantifies the efficiency of the network in sending an information between vertices as :

$$E = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}} \, .$$

- *Diameter*. The *diameter* of a network, d, corresponds to the maximum shortest path length between any pair of their nodes.

- *Betweenness distribution*. *Betweenness* of a node i, b_i, computes the proportion of shortest paths between two nodes j and k, σ_{jk}, that pass through node i, $\sigma_{jk}(i)$:

$$b_i = \sum_{j \neq i \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \, .$$

The fraction of nodes in a network with betweenness b is expressed by the *betweenness distribution*, $P(b)$. Since trees are networks without loops, for each pair of nodes there is a unique shortest path between them (Szabó et al., 2002; Bollobás and Riordan, 2004; Ghim et al., 2004).

Chapter 2

Topological characterization of phylogenies

As we have just commented on in Chapter 1, Darwin's evolutionary thought became very influential due to several reasons, such as its proposal of natural selection as the basic mechanism of evolution, its gradualist conception of evolution, and its depiction of the evolution of biodiversity with the Tree of Life. In addition, other reason why his thought became so influential was derived from his study of species distribution inside genera. In his *Origin of species*, Darwin states that species that belong to species-rich genera had more subspecific varieties, and that a vast number of species were rare (Darwin (1859), pp. 44-59). More than half a century later, in 1922, John Christopher Willis analyzed the frequency distribution of subtaxa inside taxa (e.g., species per genus). Ordering the taxa from those with the greater number of subtaxa to those with the fewest number of subtaxa, he confirmed Darwin's statement. Willis observed that few genera were species-rich, and a large amount of genera included a low number of species, referring to this frequency distribution as "hollow curve distribution" (Willis, 1922). This un-

even distribution of biodiversity aroused scientists' curiosity and, since then, a large amount of studies have been published trying to understand biodiversity distribution (Yule, 1924; Corbet, 1942; Fisher et al., 1943; Preston, 1948; Williams, 1964; Anderson, 1974, 1975; May, 1975; Zima and Horaček, 1978; Flessa and Thomas, 1985; May, 1986; Dial and Marzluff, 1989; Burlando, 1990, 1993; Tokeshi, 1993, 1996; Purvis and Hector, 2000; Hubbell, 2001; Volkov et al., 2003; Magurran and Henderson, 2003; Magurran, 2005; Pigolotti et al., 2005).

The first studies in the characterization of biodiversity distribution were carried out based on the taxa distribution in taxonomic classifications (Willis, 1922; Corbet, 1942), but the emergence of phylogenetic studies gave rise to the use of phylogenetic trees for this aim (Farris, 1973, 1976; Simberloff et al., 1981; Savage, 1983; Slowinski and Guyer, 1989; Shao and Sokal, 1990; Guyer and Slowinski, 1991; Kirkpatrick and Slatkin, 1993; Brown, 1994; Mooers and Heard, 1997; Ricklefs, 2007). Throughout this chapter, we will provide a global

[1]See Darwin (1859).
[2]See Bienaymé (1845); Galton and Watson (1874).
[3]See Willis (1922).
[4]See Yule (1924).
[5]See Corbet (1942).
[6]See Fisher et al. (1943).
[7]See Cavalli-Sforza and Edwards (1967); Harding (1971).
[8]See Sackin (1972).
[9]See Anderson (1974).
[10]See Colless (1982).
[11]See Kingman (1982b,a).
[12]See Savage (1983).
[13]See Fiala and Sokal (1985).
[14]See Dial and Marzluff (1989).
[15]See Rohlf et al. (1990).
[16]See Burlando (1990, 1993).
[17]See Kirkpatrick and Slatkin (1993).
[18]See Aldous (1995).
[19]See Mooers and Heard (1997).
[20]See Ford (2006).

1859	C. Darwin: Uneven distribution of species.[1]
1845-74	I.J. Bienaymé, F. Galton & H. W. Watson: Bienaymé-Galton-Watson branching stochastic process.[2]
1922	J. C. Willis: Hollow curve of the frequency distribution of species inside genera.[3]
1924	G. U. Yule: Earliest mathematical model of evolutionary branching.[4]
1942	A.S. Corbet: Distribution of butterflies in the Malay Peninsula.[5]
1943	R.A. Fisher: Log-series distribution to model relative species abundance.[6]
1967, 1971	L. L. Cavalli-Sforza, A. W. F. Edwards & E. F. Harding: Equal-Rates Markov model.[7]
1972	M. Sackin: Sackin's imbalance index.[8]
1974	S. Anderson: Patterns of faunal evolution.[9]
1982	D.H. Colless: Colless' imbalance index.[10]
1982	J. Kingman: Mathematical formalization of the coalescent process.[11]
1983	H.M. Savage: *The shape of evolution: systematic tree topology*.[12]
1985	Fiala: Cumulative stemminess index.[13]
1989	K.P. Dial & J.M. Marzluff: Non-random diversification within taxonomic assemblages.[14]
1990	Rohlf: Non-cumulative stemminess index.[15]
1990-93	B. Burlando: *The fractal dimension of taxonomic systems* and *The fractal geometry of evolution*.[16]
1993	M. Kirkpatrick & M. Slatkin: *Searching for evolutionary patterns in the shape of a phylogenetic tree*.[17]
1995	D.J. Aldous: Beta-splitting model.[18]
1997	A.Ø. Mooers & S.B. Heard: *Inferring evolutionary process from the phylogenetic tree shape*.[19]
2006	D. J. Ford: Alpha model.[20]

Table 2.1: Some of the main works on topological characterization and modeling of evolutionary trees.

view of some of the most important approaches developed for the inference of the evolutionary patterns through the topological characterization of phylogenetic trees,[21] with a special emphasis on the most relevant indices and models used for that aim.

2.1 Evolutionary patterns through topological characterization of phylogenies

After the first works on the inference of evolutionary patterns through the topological characterization of phylogenetic trees, the analysis of the branching patterns of binary trees[22] has become one of the main ways to characterize the patterns that evolution traces (Savage, 1983; Kirkpatrick and Slatkin, 1993; Mooers and Heard, 1997; Ricklefs, 2007). One of the most important features of the branching pattern is identified as *tree balance*, which represents the symmetry of a cladogram based on how different in sizes are the two subtrees that hang from the root of a cladogram. The balance or imbalance of a phylogenetic tree depends on variations in speciation and/or extinction rates (Kirkpatrick and Slatkin, 1993). Thus, completely balanced trees are those in which, for each bifurcation event, the new species preserve evolutionary capability of the mother species. On the opposite case, completely unbalanced trees correspond to those in which,

[21] A different approach than the one considered in this thesis, but also based on the inference of evolutionary patterns taking into account the topology of evolutionary trees, is the one emerged from the collaboration between evolutionary biology and community ecology in the last decade. This interdisciplinary approach has given rise to a large number of metrics of phylogenetic diversity for the comprehension of the evolutionary history of ecological communities (for a complete review, see (Cadotte et al., 2010)).

[22] Since most of the branching processes in a phylogenetic tree are dichotomic, the bifurcations assumed by most of the branching models are binary.

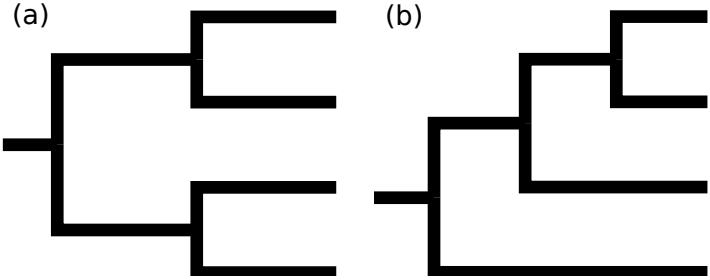


Figure 2.1: Phylogenetic tree balance. Representation of a completely balanced (a) and a completely unbalanced (b) 4-tip phylogenetic tree.

for each bifurcation event, one of the daughter species is unable to speciate, and only one species is able to speciate (Figure 2.1).

In order to explain the differences in the speciation rate, different biological foundations have been addressed, such as: ecological generalization (Rosenzweig, 1995), ecological specialization (Schluter, 1996, 2000), speciation mode (Chan and Moore, 1999), mass extinction events (Heard and Mooers, 2002), or environment effect (Davies et al., 2005).

The topological analysis of evolutionary trees provides us with two major sources of information (Moore, 2007): *topological* (branching distribution) and *temporal* (branch length distribution). Compared to those works focused on the branching patterns of cladograms, the comprehension of the topological characterization of evolutionary trees taking into account the branch length is still in its infancy (Mooers and Heard, 1997). But, ever since the first attempts on estimating the dates of the bifurcation events from paleontological evidence (Simpson, 1953), the interest on the topological characterization of phylogenetic trees through the analysis of their branch length patterns has increased (Kirkpatrick and Slatkin, 1993; Moo-

ers and Heard, 1997; Ricklefs, 2007). Given that the quantification of the branch length is extraordinarily sensitive to the reconstruction and timing methods, branch length studies have focused especially on the methodological biases derived from: phylogenetic methods (Fiala and Sokal, 1985; Rohlf et al., 1990), evolutionary models (Rohlf et al., 1990) and informative characters used for the reconstruction (Salisbury, 1999). Besides those methodological issues, the number of works focused on the biological basis of the branch length patterns, such as the effect of character evolution (Maddison, 2006; Paradis, 2008), the effect related to a punctuational or gradual evolution (Pagel et al., 2006), the correlation in branch length of functionally related genes (Li and Rodrigo, 2009), and the characterization of the frequency distribution of branch lengths in phylogenetic trees (Venditti et al., 2010), has increased in the last decade.

2.2 Evolutionary tree topological metrics

In the last decades, different topological measurement approaches have been proposed. Depending on wether they take into account the branch length or not, we can distinguish between phylogram and cladogram topological measures, respectively. In the following section we will introduce some of those measures, as well as the measures that we propose and use in this thesis.

2.2.1 Classical cladogram topological indices

These indices are exclusively based on the branching pattern of the phylogenetic trees, without taking into account the branch length. Nowadays, the number of such indices that are being used is quite large (for a review, see Kirkpatrick and Slatkin (1993), Mooers and Heard (1997) and Agapow and Purvis (2002)): I_S, $\overline{N}$, σ^2_N (Sackin,

1972), B_1, B_2 (Shao and Sokal, 1990), I_C (Colless, 1982; Heard, 1992), I' (Fusco and Cronk, 1995; Purvis and Agapow, 2002), $\sum I'$, $MeanI'_{10}$ (Agapow and Purvis, 2002), $MeanI'$ (Purvis and Agapow, 2002), C_n (McKenzie and Steel, 2000), among others. These measures constitute different ways to characterize the shape of a cladogram. From all those, the most relevant have been: Sackin's Index (I_S) (Sackin, 1972), as a depth index for cladograms, and Colless' Index (I_C) (Colless, 1982), which measures the balance of the phylogenetic trees.

Sackin's Index

The Sackin's Index, I_S, is one of the oldest measures of the shape of a cladogram (Sackin, 1972), and it adds up, over the n tips of the cladogram, the number of internal nodes, N_i, from a tip, i, to the root of the cladogram (including the root):

$$I_S = \sum_{i=1}^{n} N_i \, .$$

An equivalent definition of I_S, in terms of the number of tips, $\widetilde{N}_j$, below each internal node, j, is (Blum and François, 2005):

$$I_S = \sum_{j=1}^{n-1} \widetilde{N}_j \, .$$

Two measures derived from I_S are $\overline{N}$ and σ^2_N (Sackin, 1972; Kirkpatrick and Slatkin, 1993). $\overline{N}$ is defined as the average number of internal nodes from the tips to the root of the cladogram:

$$\overline{N} = \frac{1}{n} \sum_{i=1}^{n} N_i \, ,$$

where N_i corresponds to the number of internal nodes from tip i to the root, and n is the total number of leaves of the cladogram.

σ_N^2 is defined as the variance of the number of internal nodes from tip i to the root, N_i:

$$\sigma_N^2 = \frac{1}{n}\sum_{i=1}^{n}(N_i - \overline{N})^2 \, .$$

This measure, σ_N^2, constitutes an alternative measure for the balance of the phylogenetic trees. So, for completely balanced or symmetric, trees $\sigma_N^2 = 0$, while for completely unbalanced or asymmetric trees, the measure is maximized.

Colless' Index

Colless' Index, I_C, (Colless, 1982) computes the difference in the number of tips pending from the right branch, r_i, and from the left one, s_i, from each internal node, i, of the cladogram:

$$I_C = \frac{2}{n(n-3)+1}\sum_{i=1}^{n-1}(r_i - s_i) \, .$$

The normalizing denominator was modified later by Heard (Heard, 1992), justifying the modification to a mistake made by Colless in his original definition (Colless, 1982). So, the improved expression would be:

$$I_C = \frac{2}{(n-1)(n-2)}\sum_{i=1}^{n-1}(r_i - s_i) \, .$$

With this last normalization, this measure ranges from a value of 0 for a fully symmetric tree to a value of 1 for a fully unbalanced tree.

One of the main drawbacks of this index is that it cannot be used for trees that include polytomies.

2.2.2 Classical phylogram/chronogram topological indices

Since the middle 1980's, different tools have been developed for the topological characterization of the phylogenetic trees based on the use of evolutionary distances (character or temporal distances). A classical way to graphically represent the temporal distribution of the bifurcations is that which involves the *lineage through time* plots (Harvey et al., 1994; Nee et al., 1994b). These graphs are reconstructed retrospectively from the chronograms, plotting the logarithm of the number of ancestral lineages against time. In this kind of plots, ancestral lineages do not refer to extinct lineages, plotting only those ancestral lineages that gave rise to living descendants. This way to represent diversification rates has been extensively applied in the last decade in studies as diverse as the analysis of speciation rate of Hawaiian silverword (Baldwin and Sanderson, 1998), the effect of the habitat on speciation rates of aquatic beetles (Ribera et al., 2001), the radiation patterns of South African Restionaceae (Linder et al., 2003), and the species delimitation in the genus *Rivacindela* (Coleoptera: Cicindelidae) (Pons et al., 2006).

A huge amount of indices are being used for the topological characterization of phylograms and chronograms, such as: *stemminess* (Fiala and Sokal, 1985; Rohlf et al., 1990; Salisbury, 1999; Qiao et al., 2006), *total length, L* (Qiao et al., 2006), *whole-tree methods approach* (Chan and Moore, 2002), *gamma-statistics* (Pybus and Harvey, 2000), *sum of all the branch length, s* (Nee, 2001), *K tree score* (Soria-Carrasco et al., 2007), *branch length heritability* (Savolainen et al., 2002), *diversification rate* (Nee, 2001), *total height of cherries and sum of external branch lengths* (François and Mioland, 2007), *branch length frequency distribution* (Venditti et al., 2010), etc. In the last decades, stemmi-

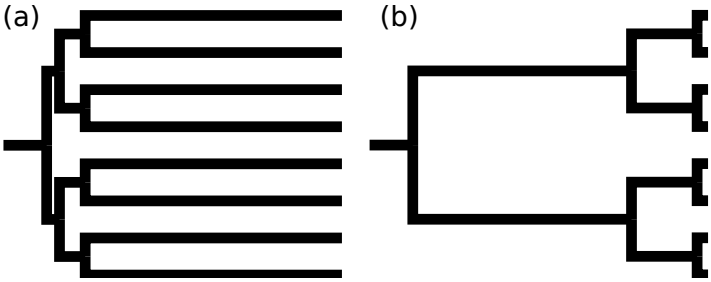


Figure 2.2: Phylogenetic tree stemminess. Representation of an 8-tip phylogenetic tree with low stemminess (a) and an 8-tip phylogenetic tree with high stemminess (b).

ness, together with all its derived measures (Fiala and Sokal, 1985; Rohlf et al., 1990; Salisbury, 1999; Qiao et al., 2006), has become one of most influential measures in the topological characterization of phylogenies based on branch length.

Stemminess

Stemminess constitutes a measure of the relative opportunity for change between clades versus change within clades. So, a stemmy tree is that with short terminal branches and long internal ones (Salisbury, 1999) (Figure 2.2).

Late in the 1970's and early in the 1980's, several works were published discussing the stemminess of phylogenetic trees (Nelson, 1979; Tateno et al., 1982). But it was in 1985 when the stemminess of a phylogenetic tree was formalized mathematically by Fiala and Sokal (1985), proposing a cumulative stemminess index, F. It was defined as the proportion of the total length of the edges of the subtree (including the length of the branch that connects this subtree with the rest of the phylogram) that is accounted for by the length of

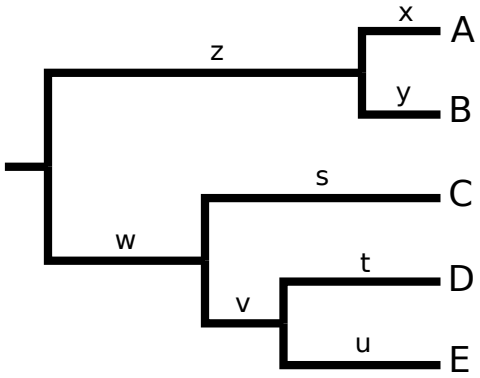


Figure 2.3: Computation of stemminess. The stemminess for subtree AB is $z/(x + y + z)$, for subtree DE $v/(t + u + v)$, for subtree CDE $w/(s + t + u + v + w)$, and the stemminess of the whole tree is the mean of these values (Fiala and Sokal, 1985).

the subtending edge of the subtree. So, based on the phylogenetic tree of Figure 2.3, the stemminess for subtree AB is $z/(x + y + z)$, for subtree DE $v/(t+u+v)$, for subtree CDE $w/(s+t+u+v+w)$, and the stemminess of the whole tree is the mean value of all these values, ignoring the length of the root branch. This cumulative stemminess index was reformulated mathematically by Rohlf et al. (1990), being defined as

$$St_C = \frac{1}{t-2} \sum_{i}^{t-2} ST_{Ci} \, ,$$

where the summation is over all internal nodes, i (excluding the root), and the stemminess value for the ith, ST_{Ci}, is

$$ST_{Ci} = \frac{W_{j \to i}}{W_{j \to i} + \sum_{k,l} W_{k \to l}} \, ,$$

where W is the branch length, and the sum is over all branches, $k \to l$, in subtrees that have i as an ancestor.

In 1990, motivated by the correlation of Fiala's stemminess index with the tree imbalance,[23] Rohlf defined a second measure for the stemminess, a non-cumulative stemminess index, R. He defined it as the mean ratio of internal branch length, in time, to the time from the branch origin to present, and is mathematically expressed as

$$St_N = \frac{1}{t-2} \sum_{i}^{t-2} ST_{Ni} \, ,$$

where t corresponds to the number of terminal taxa, and the sum is over all the internal nodes, i, excluding the root, and the stemminess value for the ith, ST_{Ni}, is

$$ST_{Ni} = \frac{W_{j \to i}}{h_j} \, ,$$

where $W_{j \to i}$ is the length of the branch between the internal node i and its ancestor, j, and h_j is the time of origin of node j.

2.2.3 Depth scaling of evolutionary trees: An allometric scaling approach

Allometric scaling relationships characterize how an observable biological quantity, such as metabolic rate or life-span, scales with the size of the biological system. In the last decade, several studies have adopted this approach for the study of transport efficiency in transportation tree-like networks, such as river basins, blood vessels or

[23] Fiala's stemminess index gives higher weights to subtrees nearer the tips of the tree.

bronchial trees, based on the characterization of the allometric scaling relationships between shape and size of those networks (West et al., 1997, 1999; Banavar et al., 1999; Garlaschelli et al., 2003; West and Brown, 2005; Makarieva et al., 2005). In that sense, we decided to apply this approach to the topological characterization of phylogenetic trees with the branch size, A, as measure of the size, and with the cumulative branch size, C, and the mean depth, d, as measures of the shape. As we will see, one of the main peculiarities of this approach is that it allows the simultaneous characterization of evolutionary tree balance[24] (shape) and evolutionary biodiversity (size).

Branch size, *A*

Based on what has been explained in Sections 1.2 and 1.3, a phylogenetic tree can be defined as a set of nodes, where each node represents a diversification event, connected by branches (links). For each node i, a subtree S_i is made up of a root at node i and all the descendant nodes stemming from this root. The subtree size, A_i, gives the number of subtaxa that diversify from node i (including itself).

Cumulative branch size, *C*

Beyond this measure of the diversity degree, A_i, the characterization of how the diversity is arranged through the phylogenies can be obtained through the cumulative branch size, C_i, a measure of the subtree shape. It is defined (Banavar et al., 1999) as the sum of the branch sizes associated to all the nodes in the subtree S_i, $C_i = \sum A_j$. For the same tree size, and restricted to binary branching events, the smallest value of the cumulative branch size is obtained for a completely symmetric, balanced tree, whereas the most asymmetric, the

[24] Unlike Colles's Index, this approach works for polytomic trees.

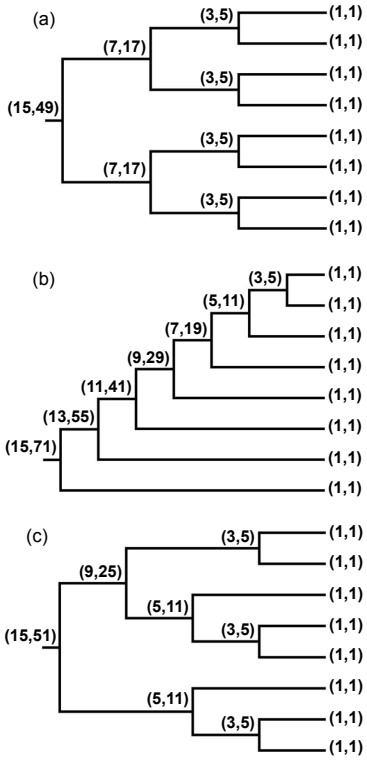


Figure 2.4: Branch size and cumulative branch size examples. The values of the subtree branch size (A) and of the cumulative branch size (C) are shown (in brackets, as (A,C)) at each node of three small example trees. (a) A completely balanced tree of 15 nodes; (b) A completely unbalanced tree of 15 nodes; (c) A subtree of 15 nodes of a real phylogenetic tree, the intraspecific *Vibrio vulnificus* phylogeny presented in full in Figure 3.4(a). Note that, for the same value of A, the value of C at the root is maximum for the fully unbalanced tree, and minimum for the balanced one.

pectinate or comb-like tree in which all branches split successively from a single one, yields the largest C_i value (Banavar et al., 1999). To be clearer, we show in Figure 2.4 the analysis of A_i and C_i for a completely balanced tree (Figure 2.4(a)) and for a completely unbalanced tree (Figure 2.4(b)). A portion of a real phylogenetic tree is also shown (Figure 2.4(c)).[25]

How the shape of the tree (i.e. the distribution of the biological diversification) changes with tree size (i.e. with the number of taxa it contains) is given by the scaling of the subtree shape, C, vs the subtree size, A, as described by the allometric scaling relation $C \sim A^{\eta}$. The symmetric tree gives $C \sim A \ln A$, which corresponds to $\eta = 1$ with a logarithmic correction, while the pectinate tree has $\eta = 2$. The natural null model for tree construction, the Equal-Rates Markov (ERM) model (Mooers and Heard, 1997; Caldarelli et al., 2004), yields a scaling $C \sim A \ln A$ equal to the symmetric tree, with $\eta = 1$.

Mean depth, *d*

The mean depth, d_i, of a subtree rooted in a node i, S_i, corresponds to the average depth of the nodes in the subtree, S_i:

$$d_i = \sum_j \frac{d_{ij}}{A_i} ,$$

where, for a given node j, the d_{ij} is its topological distance to the root of the subtree, S_i, that is, the number of nodes one has to go through so as to go from that node to the root, i (including the root in the counting), and the sum is over all nodes in the subtree, S_i. Note that here we use the mean depth over all subtree nodes, rather than just the leaves, which gives a different but related measure, $\overline{N}$ (defined above) (Sackin, 1972; Kirkpatrick and Slatkin, 1993).

[25] The Python code for the computation of the branch size, A, and the cumulative branch size, C, is included in Section H.1.

How the shape of a phylogenetic tree, i.e. the distribution of taxa diversification, changes with tree size, i.e. with the number of taxa it contains, can be analyzed with the dependence of the mean depth on subtree size $d_i = d_i(A_i)$. In the remainder, when no subindex is indicated, we understand that mean depth and other quantities refer to a whole tree or a subtree depending on the context. For a given tree size, the smallest value of the mean depth corresponds to the fully polytomic tree. The mean depth, d, as a function of tree size, A, is given in this case by

$$d_{\min} = 1 - \frac{1}{A} .$$

For large sizes the dominant order is $d_{\min} \sim 1$. The largest mean depth value for a given size is given by the fully unbalanced, or asymmetric, binary tree with a mean depth given by

$$d_{\max} = \frac{1}{4}\left(\frac{A^2 - 1}{A}\right),$$

in which for large sizes A leads to the scaling behavior $d_{\max} \sim A$. The fully balanced, or symmetric, binary tree is inside these extremes, with a mean depth given by

$$d = \frac{((A+1)\log_2(A+1) - 2A)}{A} .$$

The dominant order at large sizes is logarithmic: $d \sim \ln A$. This logarithmic scaling is not exclusive of fully balanced trees, it is also the behavior of the ERM model (Hernández-García et al., 2010), the natural null model for stochastic tree construction, in which, at each time step, one of the existing leaves of the tree is chosen at random and bifurcated into two new leaves.

The definition of the mean depth, d, is directly related to the cumulative branch size (Garlaschelli et al., 2003; Campos and de Oliveira,

2004; Camacho and Arenas, 2005; Klemm et al., 2005; Herrada et al., 2008) defined as $C = \sum_j A_j$. The sum runs over all nodes j in a tree and A_j corresponds to the size of the subtree S_j. The relationship between C and the mean depth can be obtained taking into account that the cumulative branch size can also be written as

$$C = \sum_j (d_{\text{root},j} + 1) = dA + A \ ,$$

where $d_{\text{root},j}$ is the distance of node j to the root. Thus, the mean depth of a tree is obtained as

$$d = \frac{C}{A} - 1 \ .$$

It is worth noting that the depth of a tree can also be defined taking into account only the distance from the tips to the root. This is the case of the Sackin's index, I_S, which is defined as the sum of the depths of all the leaves of the tree $I_S = \sum_j d_{\text{root},j}$ (Sackin, 1972), from which the depth measure $\overline{N} = \frac{I_S}{n}$ is constructed. Taking into account that a binary tree can be obtained as a growing tree adding at each time a speciation event we can calculate the change ΔC and ΔI_S. If the distance of the node j that speciates (leading to two new nodes) to the root is $d_{root,j}$ then

$$\Delta C = 2(d_{root,j} + 2) = 2d_{root,j} + 4 \ ,$$

while

$$\Delta I_S = -d_{root,j} + 2(d_{root,j} + 1) = d_{root,j} + 2 \ .$$

Taking into account the initial condition, that is, the root, with $C = 1$ and $I_S = 0$, one finds, for binary trees

$$C = 2I_S + 1 \,,$$

and thus, they will follow the same scaling with tree size. Also, noting that $\overline{N} = \frac{I_S}{n}$ and $A = 2n - 1$, we see that $\overline{N}$ and $d \sim \frac{C}{A}$ follow the same scaling with tree size.

Depth scaling and tree dimensionality

The different types of scaling of depth with size can be interpreted as indicating different values of the (fractal) dimensionality of the trees. This is so because $\overline{N}$, the measure defined above as the average distance from the leaves to the root, is a measure of the *mean diameter of the tree*, and because for a binary tree the total number of nodes is simply twice the number of leaves (minus one). Since the simplest definition of dimension, D, of a network (Eguíluz et al., 2003) is given by the growth of the number of nodes as the diameter increases, $n \sim \overline{N}^D$, power law scaling of the type $\overline{N} \sim n^\nu$ indicates that the tree can be thought as having a dimension $D = 1/\nu$. The logarithmic scaling in the ERM model (corresponding to $\nu = 0$ and $D = \infty$) is an example of the *small-world* behavior common to many network structures (Albert and Barabási, 2002), which is equivalent to having an effective infinite dimensionality, whereas the power law scaling reveals a finite dimension for the tree, which implies a more constrained mode of branching. The alpha model (see Section 2.3.2) produces trees with tunable dimension from 1 to ∞, and the critical activity model (see Appendix D) gives two-dimensional trees.

2.3 Modeling phylogenies

In 1924, George Udny Yule proposed the first evolutionary model for explaining the uneven distribution of subtaxa inside taxa (Yule, 1924). Since then, different evolutionary models have been proposed to improve our knowledge about those evolutionary processes that can be depicted on the evolutionary trees (Lemey et al., 2009; Hernández-García et al., 2010; Hartmann et al., 2010). Among the different modeling approaches, we can highlight two of the most important perspectives in the modeling of phylogenies. On the one hand, we find those which try to depict the evolutionary changes that take place in the genome sequence, the *substitution models*. On the other hand, we find those models that try to depict evolution through branching processes, the *branching models*.

Since the work presented in this thesis is mostly centered on the branching properties of the phylogenetic trees, the models that we have taken into account are those based on the branching approach, instead of considering the substitution models.[26] In that sense, different models have been proposed to describe evolutionary branching processes that are represented in the phylogenetic trees. As two examples, we will describe the earliest mathematical model of evolutionary branching, *Yule's model*, and one of the models proposed

[26] Within the substitution models, a distinction can be made between those models that assume that the evolutionary rate of a sequence position is constant, the so-called *homotachy models*, and those that assume that this evolutionary rate varies throughout time, the so-called *heterotachy models*. On the one hand, homotachy models are the classical ones: JC69 (Jukes and Cantor, 1969), K80 (Kimura, 1980), F81 (Felsenstein, 1981), HKY85 (Hasegawa et al., 1985), T92 (Tamura, 1992), TN93 (Tamura and Nei, 1993), etc. On the other hand, most of the heterotachy models have been devised during the last decade: covarion model (Fitch and Markowitz, 1970; Fitch, 1971), covarion-like models (Tuffley and Steel, 1998; Galtier, 2001; Penny et al., 2001; Huelsenbeck, 2002; Wang et al., 2007), mixture models (Kolaczkowski and Thornton, 2004; Lartillot and Philippe, 2004; Spencer et al., 2005; Zhou et al., 2007), covarion mixture model (Zhou et al., 2010), etc.

in the last years with an alternative depth scaling behavior to the one predicted by the Yule's model, Ford's *alpha model*. Besides, in Chapter 5 and Appendix D we propose two alternative evolutionary branching models: the *evolvability model* and the *activity model*, respectively.

2.3.1 Yule's model

It is considered the earliest mathematical model of evolutionary branching (Yule, 1924) and constitutes a special case of the birth-and-death processes, which assume that any entity existing at time t is associated to two rates:

- A *birth rate*, λ, so that during the time interval $[t, t + dt]$ a given entity has the rate $\lambda(t)dt$ of giving rise to a new entity.
- A *death rate*, $\mu(t)dt$ of dying during the time interval $[t, t + dt]$.

Yule considered the special case of λ constant and $\mu = 0$ for explaining the distribution of species inside genera. Yule's process shows that the number of species, $N(\lambda, t)$, inside a genus at time t has geometric distribution with mean $e^{\lambda t}$:[27]

$$P(N(\lambda, t) = n) = e^{-\lambda t}(1 - e^{-\lambda t})^{n-1} .$$

[27] In his original definition of the model (Yule, 1924), Yule considered an extra rate of appearance of novel genera, g. He assumed that the birth of new genera in a family of genera is the same kind of process as the birth of new species inside genera. So, replacing λ by g, the expected number of genera, N, at time t would be $N_o e^{gt}$. Historically, this rate has not been taken into account for the modeling of phylogenetic trees. So, the mentions to Yule's model on this kind of works usually refer to the Yule's model without this extra rate g.

The cladograms obtained with this model are equivalent to the ones obtained with the Equal-Rates Markov (ERM) model[28] (Cavalli-Sforza and Edwards, 1967; Harding, 1971). In ERM model, starting from a single ancestral species, one leaf is chosen at random among the tree leaves existing at the present time, and it bifurcates into two new leaves. This operation is repeated for a number of time steps or, equivalently, until the tree reaches a desired size. The topological characteristics of the constructed cladograms are surprisingly robust, being shared by apparently different models such as the coalescent model and others (Aldous, 2001). Essentially, what is needed is that different branches at a given time branch independently and with the same probabilities. When extinction is taken into account, the same topology is recovered when considering only the lineages surviving at the final time. One of the characteristics of this type of branching is a distribution of subtree sizes, A, scaling at large sizes as $P(A) \sim A^{-2}$, an outcome robustly observed in many natural and artificial systems and in classification schemes, including taxonomies (Burlando, 1990; Caldarelli et al., 2004; Capocci et al., 2008). Another important characteristic is that the mean depth of the tree, d, scales logarithmically with the number of leaves, n:

$$d \sim \log n \, .$$

It is worth noting that these results apply not only to many random branching models, but also to the simple deterministic Cayley tree, in which all internal nodes at a given level split in a fixed number of daughter nodes.

For the last decade, it has been known that real phylogenies are substantially more unbalanced than the predicted by the ERM and similar models (Aldous, 2001; Blum and François, 2006). This means that some lineages diversify much more than others, in a way that is statistically incompatible with the ERM or Yule predictions.

[28]The main difference between Yule's and ERM model is given by the fact that Yule's model is continuous in time, while ERM model works in discrete time.

The breakdown of the ERM behavior indicates that evolutionary branching should present correlations either in time or between the different branches. Mechanisms producing trees with non-ERM scaling for the depth have been identified, as for example the situation of critical branching (De Los Rios, 2001; Harris, 1963) or optimization of transport processes (Banavar et al., 1999). In the phylogenetic context, models of this type have been proposed (Aldous, 2001; Pinelis, 2003; Blum and François, 2006; Ford, 2006), although most of them lack a clear interpretation in biological terms. Among those non-ERM depth scaling behavior models, we have to highlight Ford's *alpha* model, which shows a power law scaling of the depth with the tree size (Ford, 2006).

2.3.2 Alpha model

This model is defined dynamically, that is, by a set of rules that are applied to the present state of a growing tree to find the state at the next step. At a given step in the process the tree is a set of leaves connected by terminal links to internal nodes, which are themselves connected by internal edges until reaching the root (the root itself is considered to have a single edge, which we count as internal, joining the first bifurcating internal node; with this convention a tree of n leaves has $n-1$ internal edges). Then, a probability of branching proportional to $1-\alpha$ is assigned to each leaf, and proportional to α to each internal edge. By normalization these probabilities are, respectively, $(1-\alpha)/(n-\alpha)$, and $\alpha/(n-\alpha)$. When a leaf is selected for branching, it gives birth to a couple of new ones, as in the ERM model. But when choosing an internal edge, a new leaf branches from it by the insertion in the edge of a new internal node. For $\alpha = 0$ we have the standard ERM model. For $\alpha = 1$ the completely unbalanced comb tree, in which all leaves branch successively from a main branch, is generated. Intermediate topologies are obtained for $\alpha \in (0, 1)$.

Fully polytomic tree	$C \sim A$	$d \sim 1$
Fully symmetric tree	$C \sim A \log A$	$d \sim \log A$
Fully asymmetric tree	$C \sim A^2$	$d \sim A$
Yule's model	$C \sim A \log A$	$d \sim \log A$
Ford's alpha model	$C \sim A^{\alpha+1}$	$d \sim A^{\alpha}$

Table 2.2: Cumulative branch size, C, and mean depth, d, as functions of the branch size, A, for polytomic, symmetric and asymmetric trees, and for Yule's and alpha models.

By considering the effect of the addition of new leaves on the distances between root and other nodes, Ford (2006) derived an exact recurrence relation which, when written in terms of the expected value for the average distance from the leaves to the root, $\overline{N}$, leads to:

$$\overline{N}_{n+1} = \frac{n}{n-\alpha}\overline{N}_n + \frac{2n(1-2\alpha)}{(n+1)(n-\alpha)} \, .$$

$\overline{N}_n$ is the mean depth of the leaves of a tree with n leaves. By assuming a behavior $\overline{N}_n \sim n^{\nu}$ at large n, and expanding this equation in powers of $1/n$, we get $\nu = \alpha$, so that

$$\overline{N}_n \sim n^{\alpha} \, , \text{if } \; 0 < \alpha \leq 1 \, .$$

Since, as explained above, the scaling behavior of $\overline{N}$ is the same as that of d, we have also $d \sim n^{\alpha}$. If $\alpha = 0$ the standard logarithmic scaling behavior of ERM is recovered.

The main objection to that model is the lack of a clear interpretation from the biological point of view. While the Ford model gives a simple mechanism for scaling in trees with a tunable exponent, the dynamic rule of posterior insertion of inner nodes is hard to justify

in the context of evolution (although one can think of the modeling of errors arising in phylogenetic reconstruction methods when incorrectly assigning a splitting to a non-existing ancestral species).

Chapter 3

Depth scaling in organism phylogenies

The Tree of Life is a synoptic depiction of the pathways of evolutionary differentiation between Earth life forms (Cracraft and Donoghue, 2004), and contains valuable clues on the key issue of understanding the diversification of life in the planet (Purvis and Hector, 2000). The branching pattern of the Tree of Life, which is being captured at increasing resolution by the advent of molecular tools (Rokas, 2006), can be examined to investigate fundamental questions, such as whether it follows universal rules, and at what extent random differentiation mechanisms explain the shape of phylogenetic trees. The examination of the structure of the Tree of Life can also help to infer whether evolution acts at intraspecific scales in a way different from the action of evolution at the interspecific scale. In this chapter we address these fundamental questions on the basis of a comprehensive comparative analysis of phylogenetic trees representing different fractions and domains of the Tree of Life, from intraspecific to interspecific scales. We draw from previous analysis of the geometry of the Tree of Life (Blum and François, 2006),

the characterization of other branching systems (Rodriguez-Iturbe and Rinaldo, 1997; Makarieva et al., 2005), and using tools derived from modern network theory (Garlaschelli et al., 2003; Campos and de Oliveira, 2004; Camacho and Arenas, 2005; Proulx et al., 2005; Klemm et al., 2005) to examine the scaling of the branching in the Tree of Life (LaBarbera, 1989; Webb et al., 2002). Our analysis is based on the comparison of the results derived from the analysis of inter- and intraspecific phylogenetic trees, to test for the preservation of branching patterns across evolutionary scales, and against those derived from the analysis of randomly-generated trees to test whether the depth scaling derived can be modeled using simple, random branching rules.[1]

3.1 Materials and methods

3.1.1 Phylogenies databases

On June 30th 2007 we downloaded the 5,212 phylogenetic trees available at that time in the database TreeBASE (2010). TreeBASE constitutes a large database of interspecific phylogenies, which were collected from previously published research papers. The size of trees oscillates from 10 to 600 tips. Most of the bifurcations in these trees are binary, as confirmed by the fact that the ratio between the number of tips and the total number of nodes is 0.52 when averaged over all the trees (for perfect binary trees, the ratio is 0.50).

As a comprehensive database comparable to TreeBASE does not exist for intraspecific phylogenies, we constructed an intraspecific dataset by manually compiling 67 intraspecific phylogenies from several published phylogenetic analysis (see Table A.1). We compiled this

[1] The work presented in this chapter has been published in Herrada et al. (2008).

	INTRA	INTER
Animalia	24	26
Archaea	0	3
Bacteria	18	9
Fungi	6	13
Plantae	6	8
Protozoa	4	6
Viruses	9	2
TOTAL	67	67

Table 3.1: Break-down of the number of analyzed intra- and interspecies trees with respect to taxa.

dataset in such a way that it contains: 1) Organisms from the main different environments (terrestrial, marine and fresh water), climatic regions (from polar to desert), and branches of life (Table 3.1). 2) Phylogenetic trees reconstructed with three of the main phylogenetic tree estimation methods: neighbor-joining, maximum parsimony and maximum likelihood.

In order to test whether the results derived from the examination of the relatively small (67 phylogenies) intraspecific data base can be compared with the much larger (5,212 phylogenies) set of interspecific phylogenies extracted from TreeBASE, we sampled the literature to construct a dataset of 67 interspecific phylogenies drawn from the literature (see Table A.2) using the same criteria as those to derived the intraspecific phylogeny data base (see Table A.1). The intra- and interspecific phylogenies derived from the literature ranged between 30 and 170 tips, and they contained mainly binary branching events. An example for each kind of phylogenies is shown in Figures 3.4(a) and 3.4(b).[2]

[2]Given that for the computation of the branch size, A, and the cumulative branch size, C, the input of the phylogenetic trees has to be defined in columns format,

3.1.2 Branch size and cumulative branch size distributions

We associate two quantities to each node i of a phylogenetic tree, the size (number of nodes), A_i, of the subtree, S_i, made up of node i and all the descendant nodes below it, that is, the subtree which does not contain the global root of the original tree, and the cumulative branch size, C_i, defined as the sum of the branch sizes associated to all the nodes in the subtree S_i, $C_i = \sum A_j$. To characterize the probability distributions of the A_i and C_i values on a particular phylogenetic tree we compute the respective complementary cumulative distribution functions (CCDF): $F(A)$ = probability($A_i > A$), and $F(C)$ = probability($C_i > C$). We observe that these quantities scale, for large values of A and C, as power laws: $F(A) \sim A^{1-\tau_A}$ and $F(C) \sim C^{1-\tau_C}$. The exponents τ_A and τ_C, thus, characterize the probabilities of $\{A_i\}$ and $\{C_i\}$: $P(A) \sim A^{-\tau_A}$ and $P(C) \sim C^{-\tau_C}$, respectively.

3.1.3 Allometric scaling relationship

We observe that a functional relationship among the values of C and A, i.e. among shape and size, exists and can be fitted by a power law, $C \sim A^{\eta}$, characterized by an exponent η. Since this relationship encodes the variation of a system property as size is varied, we can call this an allometric scaling relationship, so as to stress its connections with other functional relationships relating function and size (LaBarbera, 1989; Banavar et al., 1999; Brown et al., 2004). We note that introduction of the change of variables $C \sim A^{\eta}$ into $F(C) \sim C^{1-\tau_C}$ leads to $F(C) \sim A^{\eta(1-\tau_C)}$, from which $\eta = \frac{1-\tau_A}{1-\tau_C}$. Thus, only two out of the three exponents are independent. As simple examples for which the above exponents can be computed by direct counting, we mention the pectinate or fully unbalanced

instead of Newick format (the convention commonly used to write out phylogenetic trees in a text file), in Section H.2 we included the Python code for the conversion from Newick to columns format.

tree, i.e. a tree in which all branching occurs successively along a single branch, characterized by the exponents $\tau_A = 0$, $\tau_C = 1/2$, $\eta = 2$, or the fully symmetric or Cayley tree, characterized by $\tau_A = 2$, and $C \sim A \ln A$, which except for the weak logarithmic correction corresponds to $\eta = 1$ and $\tau_C = 2$.

In order to investigate whether observations differ from random expectations, we have compared the allometric scaling found here with the prediction of a null model (Harvey et al., 1983), the Equal-Rates Markov (ERM) model. The ERM model was attributed to Harding (1971), and to Cavalli-Sforza and Edwards (Cavalli-Sforza and Edwards, 1967), although it is based on models of the diversification process that date back at least to Yule (1924). The main assumption of the ERM model is that the phylogeny is the product of random branching. This is the result when the "effective speciation rate" (the difference between extinction and speciation rate) is equal for all species. The effective speciation rate may change chronologically, provided that it is the same for all lineages at a given time (Yule, 1924). For this model we obtain $C \sim A \ln A$, or $\eta = 1$, and also $\tau_A = \tau_C = 2$. The random asymmetries introduced by the ERM are not strong enough to change the scaling behavior from the symmetric tree result.

3.2 Results

The branch-size CCDF displays power-law tails of the form $F(A) \sim A^{1-\tau_A}$ for large branch size, A, (Figure 3.1(a)). The power-law exponents τ_A are remarkably similar for the datasets analyzed: $\tau_A = 1.76 \pm 0.03$, and 1.74 ± 0.02 for intra- and interspecific phylogenies, respectively. Similarly, the cumulative-branch-size CCDF also displays a power-law tail of the form $F(C) \sim C^{1-\tau_C}$ at large C, with a similar agreement between the exponents of the intra- and interspe-

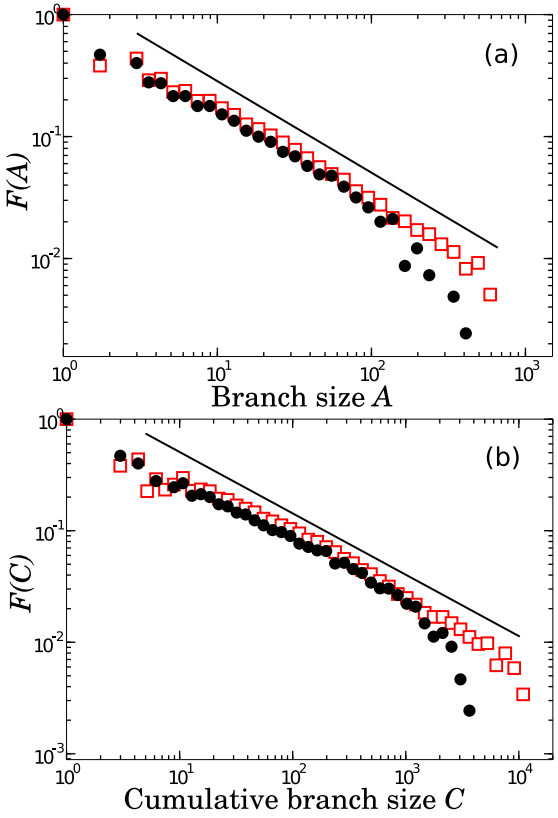


Figure 3.1: Intra- and interspecific average distributions. Cumulative complementary distribution functions (CCDFs) averaged and logarithmically binned over all phylogenetic trees in the intraspecific (black solid circles) and interspecific (red empty squares) datasets. (a) CCDF of branch size, $F(A)$. Solid line corresponds to a power law $F(A) \sim A^{1-\tau_A}$ with the exponent given by the best fit to the interspecific dataset $\tau_A = 1.74$. (b) CCDF of the cumulative branch size, $F(C)$. The line corresponds to a power law with the exponent given by the best fit to the interspecific dataset $\tau_C = 1.53$.

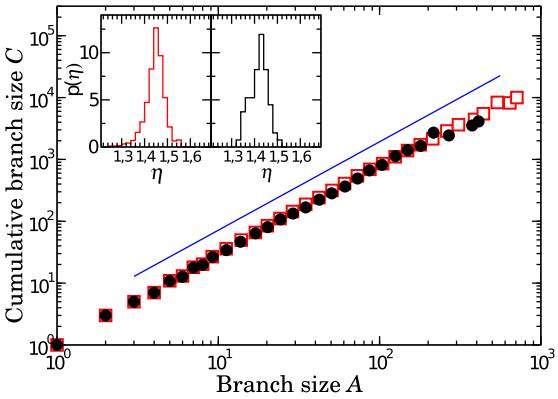


Figure 3.2: Intra- and interspecific allometric scaling. Plot of the logarithmically binned set of values of branch size, A, vs cumulative branch size, C, for the intraspecific (black solid circles) and interspecific (red empty squares) datasets considered. The line corresponds to a power law $C \sim A^{\eta}$, with the exponent given by the best fit through all data, $\eta = 1.44$. The inset shows probability distributions of the values of η fitted to each individual tree (left: interspecific, right: intraspecific datasets) illustrating the small dispersion in the values.

cific datasets: $\tau_C = 1.53 \pm 0.02$ and 1.53 ± 0.02, respectively (Figure 3.1(b)). The discrepancy observed between the two datasets at the tail of the distributions can be explained by the different sizes of the typical trees on them: each tree contributes a natural cutoff to the overall distribution, and since the intraspecific trees are smaller in average, their cutoff appears at smaller tree sizes.

The allometric exponent, η, that characterizes the scaling of tree shape with tree size (Figure 3.2), is also remarkably similar for the intraspecific ($\eta = 1.43 \pm 0.01$) and the interspecific ($\eta = 1.44 \pm 0.01$) phylogenies. This constancy of the exponents is still more remark-

able when realizing (inset of Figure 3.2) that it does not only apply to average properties of sets of intraspecific and interspecific trees, but also to individual phylogenies of groups of organisms pertaining to different kingdoms and living across widely contrasting environments, as it is reflected by the very narrow range of η ($\langle\eta\rangle = 1.47$, $\sigma = 0.03$, Figure 3.2).

The scaling exponents for our large interspecific dataset are also matched almost perfectly (Figure 3.3) by those derived from a set of 67 interspecific phylogenies randomly drawn from the published literature (see Appendix A), thereby validating the uniformity of the scaling rules of the broad interspecific phylogenies and the smaller set of intraspecific ones used here. The later was also derived from a similar random sample taken from the published literature (see Appendix A). We see that, despite their different size, the two interspecific datasets display the same behavior. Any bias in the manual selection procedure with respect to TreeBASE, if present, is weak enough to have no impact on the topological scaling behavior. In addition, there is perfect agreement between the scaling of the three datasets, except for the largest tree sizes for which there is poor statistics in the smaller datasets. This gives further support to the universality of the scaling found. As examples, we illustrate in Figure 3.4 the tree structures and the allometric scaling for an intraspecific (Figure 3.4(a) and (c)) and an interspecific tree Figure 3.4(b) and (d)), respectively.

The allometric scaling of $C \sim A^{1.44}$ derived from our analysis falls somehow in between those obtained by simulated phylogenies derived from two extreme topologies: The symmetric tree gives $C \sim A \ln A$, which corresponds to $\eta = 1$ with a logarithmic correction, while the pectinate tree has $\eta = 2$. The natural null model for tree construction, the ERM model (Mooers and Heard, 1997; Caldarelli et al., 2004), yields a scaling $C \sim A \ln A$ similar to the symmetric tree with $\eta = 1$ but different from the scaling displayed by empirical inter- and intraspecific phylogenies, particularly for large ones (see Figure 3.5). Therefore some topological aspects of phylogenetic

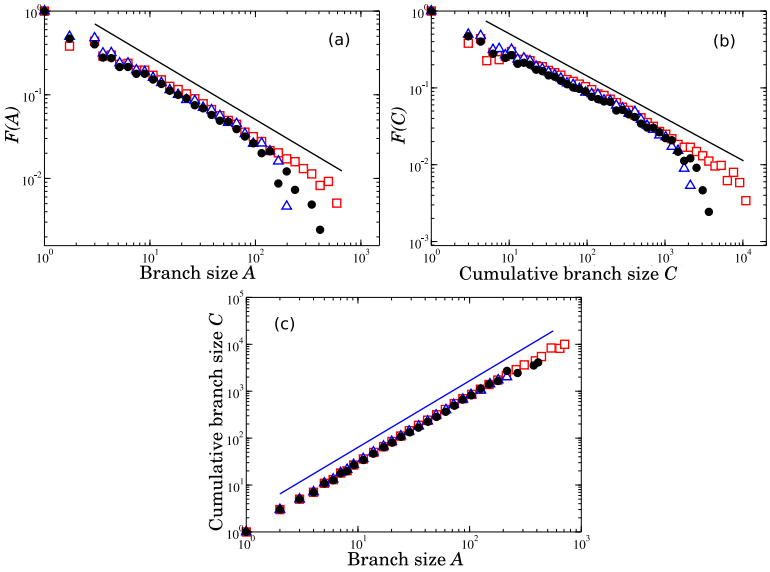


Figure 3.3: Inter- and intraspecific scalings. Same as Figures 3.1 and 3.2 but adding the manually compiled interspecific data. Cumulative complementary distribution functions (CCDFs) for branch size ($F(A)$ (a) and cumulative branch size ($F(C)$ (b), and the allometric scaling relation ($C \sim A^{\eta}$) (c) averaged and logarithmically binned over all phylogenetic trees. Red empty squares are for the interspecific TreeBASE dataset, black solid circles are for the manually compiled intraspecific dataset, and blue empty triangles are for the new manually compiled interspecific dataset of reduced size. Solid lines are power laws fitted to the TreeBASE behavior (red empty squares).

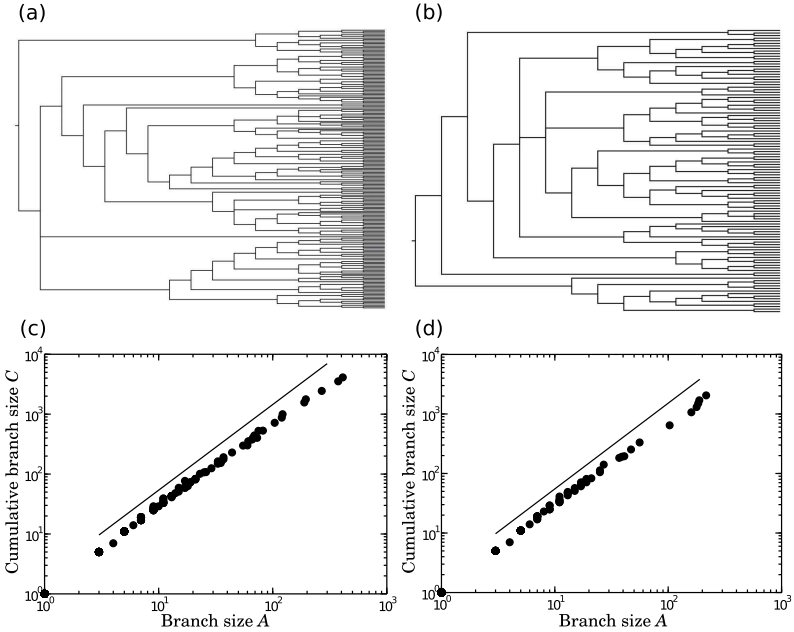


Figure 3.4: Examples of intra- and interspecific phylogenetic trees. (a) An example of an intraspecific phylogenetic tree: different strains of the bacteria *Vibrio vulnificus* (Lin et al., 2003). Most of the branchings are binary, but there are some 3rd order branchings. (b) An example of an interspecific phylogenetic tree: the catfish species (order Siluriformes) (Sullivan et al., 2006). Most of the branchings are binary, but there are some 3rd order branchings. (c) The allometric scaling plot showing the relationship of cumulative branch size (C) to branch size (A) from each node of that tree. The solid line corresponds to the fitting $C \sim A^{1.43}$ to this intraspecific dataset. (d) The allometric scaling plot showing the relationship of cumulative branch size (C) to branch size (A) from each node of that tree. The solid line corresponds to the fitting $C \sim A^{1.44}$ to this interspecific dataset.

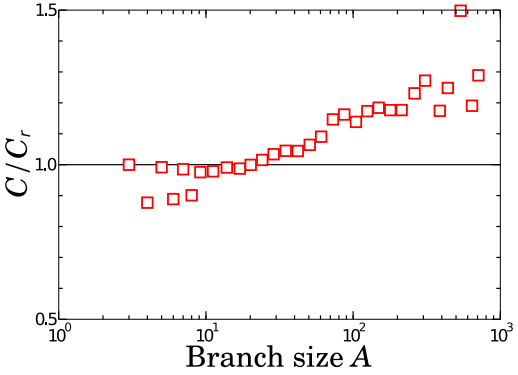


Figure 3.5: Allometric scaling (random). Plot of the logarithmically binned set of values of C as a function of A for the interspecific data (red empty squares), normalized by the prediction from the ERM model (the horizontal line). Data systematically deviate from ERM, especially for large size A.

trees are not adequately reproduced by the ERM model. Our results imply that successful lineages diversify more profusely than expected under random branching, generating the large imbalances that characterize emerging depictions of the Tree of Life (Blum and François, 2006). Alternative models introducing correlations, such as the proportional-to-distinguishable-arrangements (PDA) model (Pinelis, 2003; Blum and François, 2006) or the beta splitting model (Aldous, 2001), could generate more realistic phylogenies. Guided by previous biological allometric scaling analysis, we have assumed a power-law scaling of the form $C \sim A^{\eta}$. However, as we will see in Chapters 4 and 5, other ansatz could also fit the data. The important point, however, is that these modeling approaches should give similar scaling properties for intra- as for interspecific branching.

3.3

Discussion

Traditionally, microevolutionary and macroevolutionary processes have been studied independently by population geneticists and evolutionary biologists, respectively (Simons, 2002). The divide between these two levels of generation of biological diversity is an old one, rooted in the controversy between Darwinian gradualism and the saltationism proposed by others, prominently paleontologists, to explain macroevolutionary processes (Mayr, 1982). The debate as to whether macroevolution is wider process than the mere accumulation of microevolutionary events remains active (Simons, 2002; Grantham, 2007; Erwin, 2000), although refined paleontological evidence supports the continuum between micro- and macroevolution for some lineages (Kutschera and Niklas, 2004). The results presented here show that the branching and scaling patterns in intraspecific and interspecific phylogenies do not differ significantly for the topological properties we have calculated. Thus, shall saltation processes be a factor at the macroevolutionary level? This is not reflected in the topology of phylogenetic branching as examined here. Evidence for possible differences in phylogenetic topologies between the inter- and intraspecific levels may require a detailed analysis of branching times, which we have not attempted in this chapter.

Processes leading to scaling laws in size distributions in natural systems have been formulated as growth models (Yule, 1924; Simon, 1995). Many of the findings carry over to scaling properties found in networks (Bornholdt and Ebel, 2001) and their description in terms of branching processes (Durrett, 2007). But most of these models predict branching topologies similar to the ERM model. An alternative approach to the understanding of the observed exponent would be to trace analogies with scaling laws in different branching systems (Rodriguez-Iturbe and Rinaldo, 1997; Makarieva et al., 2005; Brown

et al., 2004), which have been explained by invoking a natural optimization criterion based in the fact that the observed trees contain the largest possible number of apices within the smallest number of branching levels. For binary trees of size A, where nodes are restricted to occupy uniformly a D dimensional Euclidean space, the minimum value of C scales as A^{η}, with $\eta = \frac{(D+1)}{D}$. This scaling also describes the D-dimensional tree with the maximum size for a given depth (the average distance between root and leaves). The value of η obtained in our phylogeny analysis, $\eta \cong 1.44$, is achieved only for optimal trees restricted to spaces of $D \cong 2.27$ dimensions. Given the apparently unlimited number of variables that may yield differences among taxa, restricting their representation to a space with such a small number of dimensions seems unreasonable. This interpretation suggests that the evolutionary process yielding the observed phylogenies is not the most parsimonious one, which could potentially yield a similar biodiversity with fewer branching levels. In fact, the natural choice $D = \infty$ gives an optimal exponent $\eta = 1$, which corresponds to the ERM value and departs from observed scaling. Optimal traffic networks (Barthélemy and Flammini, 2006) also led to the exponent $\tau_A = 2$ which departs from the empirical scaling exponent reported here for phylogenetic trees.

Besides the ERM-like scaling described by most of the proposed evolutionary branching models (Yule, 1924; Cavalli-Sforza and Edwards, 1967; Harding, 1971) and the power law scaling proposed here and depicted by the model defined by Ford (2006) (in Appendix D we propose the *activity model*, which displays a power law scaling), two different alternative scaling behaviors have been proposed. On the one hand, the AB model (Aldous, 2001), which leads to a squared logarithmic scaling behavior and also gives a reasonable fitting to our data. On the other hand, more recently, Stich and Manrubia (2009) suggest that the non-ERM behavior depicted by the real phylogenies is a small-size transient behavior, which would cross-over to the ERM scaling as larger tree sizes become available.

In summary, the remarkably similar allometric exponents reported here to characterize universally the scaling properties of intra- and interspecific phylogenies across kingdoms, reproductive strategies and environments, strongly suggests the conservation of branching rules, and hence of the evolutionary processes that drive biological diversification, across the entire history of life.[3] Although at short branch sizes the topology of observed phylogenies cannot differ much from that expected under random and symmetric trees, due to the restriction of binary bifurcations in phylogenetic tree reconstruction, significant departures become universally evident as trees become larger, where the null ERM model and real phylogenies differ (see Figure 3.5). These deviations suggest (a) that the evolution of life leads to less biodiversity than an optimal tree can possibly generate; and (b) the operation of a mechanism generating a correlated branching, where some memory of past evolutionary events is maintained along each branch. This correlated branching pattern implies that entities that diversify faster than average lead to new biological forms that, in turn, diversify more than average. Invariance across the broad scales considered here indicates that relatively simple rules govern the phylogenetic branching and the unfolding of biodiversity. Their deviation from random models indicates that evolutionary success is a correlated trait within lineages, yielding present asymmetries in the structure of the Tree of Life.

[3] After having published those results that we present in this chapter (Herrada et al., 2008), a paper was published pointing to the non-universality of our result given by the effect of the outgroup in the allometric scaling of the phylogenetic trees (Altaba, 2009). In Appendix B we confirm our results presented in this chapter.

Chapter 4

Depth scaling in gene phylogenies

Together with the results presented in Chapter 3, several analysis of the topological properties of phylogenies have shown universal patterns of phylogenetic differentiation (Dial and Marzluff, 1989; Burlando, 1990, 1993; Blum and François, 2006). This means that the impact of evolutionary forces on the shape of phylogenetic trees is, at least as the different quantifiers studied captured it, similar across a broad range of scales, shaping the diversity of life on Earth, from macro-evolution to speciation and population differentiation, and across diverse organisms such as eukaryotes, eubacteria, archaea or viruses. This, and the fact that evolutionary forces work at molecular level, motivates the study of the topology of evolutionary relationships among molecular entities, looking for patterns of differentiation at such microscopic level. Taking into account that, as we introduced in Section 1.2.5, a group of evolutionary related genes is considered a *gene family*, in the present chapter we carry out an analysis of the topological properties of gene family phylogenies by using the depth scaling approach described in Chapter 2.

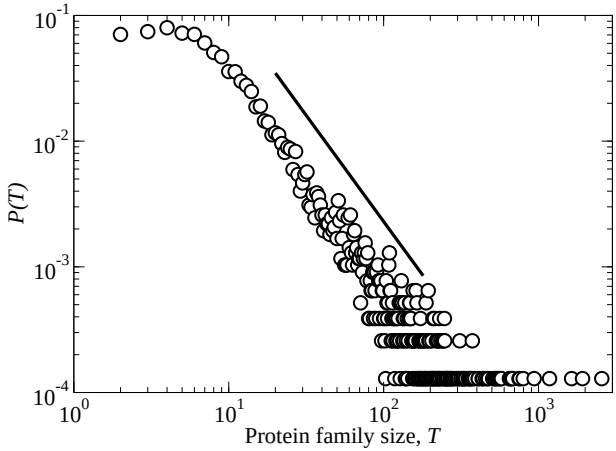


Figure 4.1: Protein family size distribution. Distribution of the size of the PANDIT protein families. Black line corresponds to a power-law $P(T) \sim T^{-\gamma}$, with a fitted exponent $\gamma = 1.69 \pm 0.05$.

4.1 Datasets

We have analyzed the 7,738 protein families[1] available on May 27th 2008 in the PANDIT database (`http://www.ebi.ac.uk/goldman-srv/pandit/`) (Whelan et al., 2003, 2006). PANDIT is based upon Pfam (`http://pfam.sanger.ac.uk/`) (Bateman et al., 2004), and constitutes a large collection of protein family phylogenies from different signalling pathways, cellular organelles and biological functions. The size of each of the phylogenies, T, ranges from 3 to more than 2000 tips and, agreeing with previous reports (Huynen and van

[1]Note that, from Section 1.2.5, the term *protein family* is referred to a gene family of genes that encode for proteins, but both are often used as nearly synonymous.

Nimwegen, 1998; Harrison and Gerstein, 2002; Koonin et al., 2002; Luscombe et al., 2002; Unger et al., 2003), it is distributed according to a power-law distribution $P(T) \sim T^{-\gamma}$ (see Figure 4.1). Most of the bifurcations in these phylogenies are binary, with only 22% of polytomic bifurcations.

For the comparative analysis between protein phylogenies and species phylogenies, we used the set of 5,212 interspecific phylogenetic trees from TreeBASE database used in Chapter 3, downloaded on June 30th 2007.

4.2 Results

The analysis of the 7,738 protein phylogenies of PANDIT database shows that the scaling of the mean depth with tree size lies between the two extreme topologies for binary trees (fully unbalanced and fully balanced trees). The raw data is not scattered between the extreme cases but instead it is concentrated around some intermediate behavior that depends on the size of the trees. In order to characterize the dependence of the depth with tree size we have logarithmically binned the data. In Figure 4.2 we show the dependence of the square root of the depth for different tree sizes which suggests a square logarithmic scaling of the form $d \sim (\ln A)^2$, while the fully unbalanced tree shows a linear dependence $d \sim A$, and the fully balanced tree shows a logarithmic dependence of the form $d \sim \log A$. We note that a power-law, $d \sim A^{0.44}$, as discussed in Chapter 3, is also good, but a square logarithmic scaling seems to fit better.

With the aim of determining if the average mean depth scaling described for the whole PANDIT database is also preserved after discriminating the different protein functions, we decided to analyze

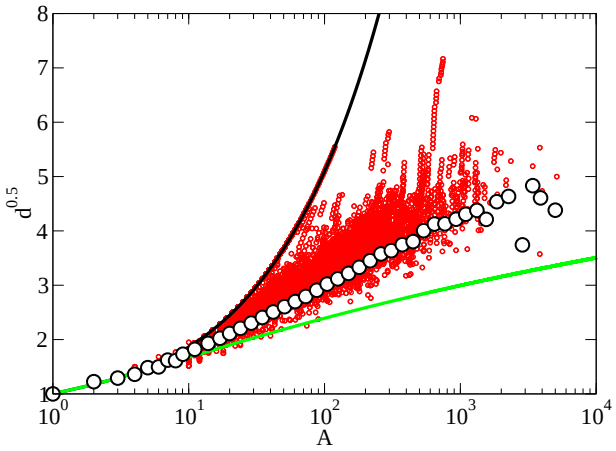


Figure 4.2: Depth scaling of protein phylogenies. Mean depth scaling for all protein families in the PANDIT database (red dots, where each dot represents a subtree) and the corresponding averaged binned depth (black empty circles). The black and green lines correspond to the two extreme topologies for binary trees, fully unbalanced and fully balanced trees, respectively.

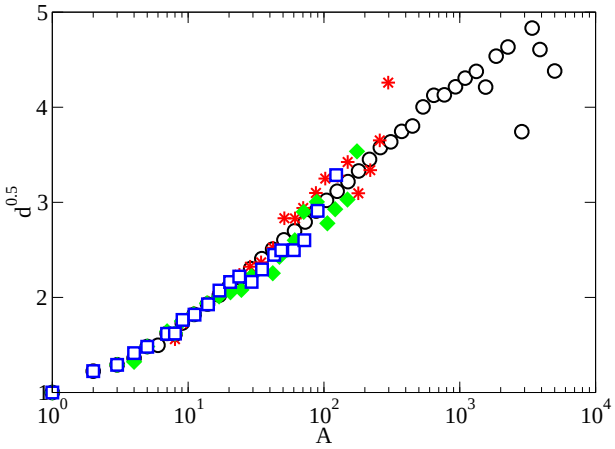


Figure 4.3: Depth scaling of different protein functions. Binned values of the mean depth for nuclear (blue empty squares), structural (green solid diamonds) and metabolic (red stars) protein families. The black empty circles represent the averaged binned depth for the whole PANDIT database.

the scaling of the mean depth as function of the tree size for different protein functions (nuclear, structural, metabolic, etc.). The results obtained are summarized in Figure 4.3. There it is observed that the depth of different protein functions shows the same behavior as the one described by the whole PANDIT dataset, without any significant deviation from the averaged mean depth scaling for the whole PANDIT database (Figure 4.2). This result suggests the universality of the depth scaling of protein phylogenies.

This generality of the depth scaling behavior observed for protein phylogenies is even more remarkable when protein phylogenies are compared with the species phylogenies obtained from the TreeBASE database (Figure 4.4). The comparative analysis between the mean depth scaling of PANDIT and TreeBASE shows a similar scaling of

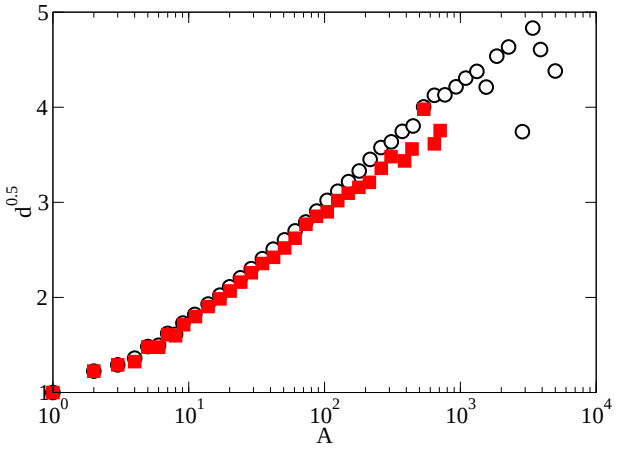


Figure 4.4: Protein vs organism phylogenies. Averaged and binned mean depth for organisms in TreeBASE (red solid squares) and for protein phylogenies in PANDIT (black empty circles).

the mean depth with the tree size for both datasets. Although in the previous work described in Chapter 3 with organism phylogenies the depth scaling was fitted to a power-law (Herrada et al., 2008), the squared logarithmic scaling $d \sim (\ln A)^2$ shows a slightly better fitting with the protein families. In order to elucidate which scaling provides a better description of the scaling of the mean depth one should consider larger trees which are not available at the moment. However, the analysis of protein phylogenies shows that the trees follow some universal mechanism as they speciate and that this mechanism seems to be the related to the speciation at the species level.

There is some dispersion of the mean depth for the whole PANDIT dataset observed in Figure 4.2, and this is due to unbalanced bifurcations in some specific trees. This increase in the presence of

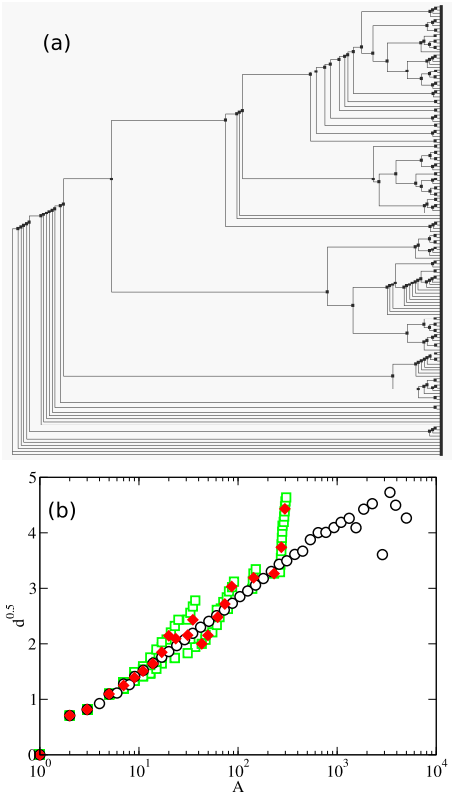


Figure 4.5: Example of the mean depth behavior in a specific phylogenetic tree. (a) Phylogenetic tree corresponding to the Probable molybdopterin binding domain family (PF00994), with a high presence of unbalanced bifurcations close to the root. (b) Mean depth scaling of Probable molybdopterin binding domain family phylogenetic tree, where the green empty squares correspond to the protein family and red solid diamonds represent to the corresponding averaged and binned data. Black empty circles represent the averaged and binned set for all the protein families of PANDIT.

unbalanced bifurcations is reflected as a fast increase, characteristic of fully unbalanced trees. These regions with a high number of unbalanced bifurcations are most of the times close to the root, which can be related with a lack of resolution in the reconstruction process. In Figure 4.5 we show a detailed example of a phylogenetic tree with a region with a high presence of imbalance in the bifurcations close to the root, that leads to a dispersion from the mean depth scaling in the range $A \in \left(2 \times 10^2, 3 \times 10^2\right)$, preserving the previously described universal mean depth scaling behavior in most of the size range, from 1 to 2×10^2. The fact that the dispersions from the mean are restricted only to local behaviors inside some regions of the phylogenetic trees leads us to highlight the universality of the average depth scaling behavior found in the protein phylogenies from PANDIT database.

4.3 Discussion

The increase of the high-throughput "-omics" studies has fueled the historical debate about how the gene-level evolution shapes the species-level evolution (Morris, 2000; Carroll, 2005; Roth et al., 2007). This debate connects with the one of the (dis)continuity between micro- and macroevolution, or gradualism versus saltationism (Erwin, 2000; Simons, 2002; Grantham, 2007). Following this controversy, if the universality of the scaling properties of the intra- and interspecific shown in the previous Chapter 3 suggested the conservation of the evolutionary processes that drive biological diversification across the entire history of life, in the present chapter we showed that the universality of the scaling properties is also extrapolable to the gene-level. The results presented here show that the branching and scaling patterns in protein families do not differ significantly from the patterns observed in species phylogenies, at

least for the topological properties we have calculated. We do not observe any discrepancy between the shape of protein phylogenies and species phylogenies, therefore evidence for possible differences in phylogenetic trees among protein families with different biological functions, as well as differences in phylogenetic trees between the gene and species-level, may require a detailed analysis of branching times.[2]

In 2006, Cotton and Page published a comparative analysis between human gene phylogenies and different species phylogenies (Cotton and Page, 2006). They found quantitative differences between human paralogous gene and orthologous gene phylogenies. Our results seem to differ from the one described by Cotton and Page. However that work focused on the comparison between paralogous and orthologous gene families, while we analyzed complete protein families, which included both paralogous and orthologous protein members, focusing our analysis on the comparison between protein and species phylogenies, instead of between paralogous and orthologous gene phylogenies. Otherwise, our approach is based on a qualitative analysis, while Cotton's approach is based on a quantitative analysis. This implies that, despite finding quantitative differences between paralogous and orthologous gene phylogenies, we would expect that, from a qualitative point of view, both phylogenies would display similar behavior to the one that we have just described for complete protein phylogenies and organism phylogenies.

Summing up, the universal scaling properties at gene- and species-level, characterized by the similar scaling described, strongly suggest the universality of branching rules, and hence of the evolutionary processes that drive biological diversification across the entire history of life, from genes to species.[3]

[2] A characterization of the distribution of the branching times all over the Tree of Life is provided in Chapter 7.

[3] In Appendix C we provide an attempt to depict the extent to which speciation and gene duplication events, i.e. the evolutionary processes responsible for the two

major forms of homology (orthology and paralogy, see Section 1.1.2), contribute to protein family diversification.

Chapter 5

Depth scaling modeling

The depth scaling behavior shared by protein and species phylogenies (described in Chapters 3 and 4) can be explained by different branching mechanisms. In this direction, during the last decade, several models have been published proposing different mechanisms to capture the topology of phylogenetic trees (Aldous, 2001; Pinelis, 2003; Blum and François, 2006; Ford, 2006; Stich and Manrubia, 2009; Hernández-García et al., 2010). Most of the models proposed give a logarithmic scaling of the mean depth, i.e. ERM-type for large sizes (Yule, 1924; Cavalli-Sforza and Edwards, 1967; Harding, 1971); the AB model proposed by Aldous (2001) is one of the few models that deviate from the ERM-like scaling leading to a squared logarithmic $d \sim (\ln A)^2$ (see also Blum and François (2006)); models with power law scaling of the mean depth $d \sim A^{\eta}$ have also been identified in statistical terms (Ford, 2006) and in terms of (simplified) evolutionary events (in the sense specified by Pinelis (2003)) (see the *activity* model described in Appendix D). An alternative explanation of the scaling properties of the phylogenetic trees (Stich and Manrubia, 2009) suggests that the non-ERM behavior is a small-size transient behavior, which would cross-over to the ERM scaling $d \sim \ln A$ as larger tree sizes become available.

The essential ingredient to obtain non-ERM behavior in tree scaling is the presence of temporal correlations, which leads to asymptotic or just finite-size deviations with respect to the ERM behavior depending on whether these correlations are permanent or restricted to finite but large times. In this perspective we introduce here a simple model which aims at explicitly introducing the presence of correlations. It is based on the concept of *evolvability*, i.e. the ability to evolve (Dawkins, 1989; Brookfield, 2009), as a biological characteristic which is itself inherited by sister species in speciation events. At each branching event, each species gives rise to two new species. For these two daughter species, we allow two possible outcomes.

- with probability p, the new species inherit the evolvability of the mother species, i.e. they have the same capacity as the mother species to speciate again;
- with probability $1 - p$, one of the daughter species is unable to speciate again, that is, only one of the two daughter species preserves the ability to evolve. Stemming from the definition of *robustness* as the property of a system to remain invariant in the presence of genetic or environmental perturbations (Masel and Siegal, 2009), we consider a species' inability to speciate its robustness.

The first case gives rise to a symmetric speciation event, in which the two species emerging from the speciation event are similar, while the second one gives rise to asymmetries in the tree. If $p = 1$, we recover the completely balanced binary tree, while in the other extreme $p = 0$, the topology obtained is the completely unbalanced binary tree (Figure 5.1). Thus the model combines symmetric with asymmetric branching introducing correlations (since one occurrence of the asymmetric event precludes further speciation on that branch), with the proportion determined by the parameter p. [1]

[1] The Python code that we used to simulate the model is included in Section H.3.

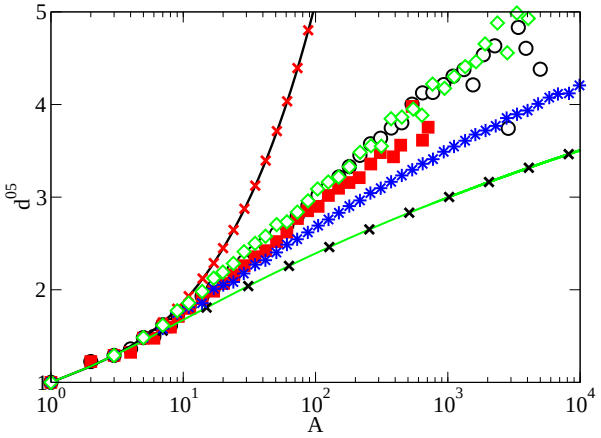


Figure 5.1: Depth scaling of the evolvability model. The mean depth scaling of the trees generated for $p = 1$ (black crosses) and for $p = 0$ (red crosses) coincides with the mean depth scaling of the fully balanced (green line) and unbalanced binary trees (black line), respectively. The trees for $p = 0.25$ (green empty diamonds) adjust the average behavior of protein (black empty circles) and organism datasets (red solid squares) very well. The blue stars correspond to trees for $p = 0.5$.

We have generated trees with this algorithm and observed that, by choosing $p = 0.25$, the depth scaling of the trees is very close to the one observed for phylogenetic trees in both PANDIT and TreeBASE (Figure 5.1). This result identifies the prevalence of unbalanced branching events (occurring with frequency $1 - p = 0.75$) with respect to balanced ones ($p = 0.25$), which is consistent with results described in previous works (Mooers and Heard, 1997; Aldous, 2001; Blum and François, 2006).

It should be said, however, that the correlations introduced by our model are not permanent and finally a crossover to the random behavior appears for long sizes. To see this, we calculate the analytical expression of the average depth, d. Taking into account that the expected number of offsprings of a pair of sister nodes is $2z = 4p + 2(1-p) = 2(1+p)$, starting with the root, the expected number of nodes after n branching events is

$$\langle A \rangle = 1 + 2\Big[\sum_{i=0}^{n-1} z^i\Big] = 1 + 2\frac{z^n - 1}{z - 1}\,,$$

where $z = 1 + p$ is the expected offspring per sister node. The expected value of the cumulative branch size is given by

$$\langle C \rangle = 1 + 2\Big[\sum_{i=0}^{n-1} z^i(i+2)\Big] = 1 + 2\Big\{z\frac{(n-1)z^n - nz^{n-1} + 1}{(z-1)^2} + 2\frac{z^n - 1}{z - 1}\Big\}\,.$$

We have at large n that $\langle A \rangle \sim z^n$ and $\langle C \rangle \sim nz^n$. Taking into account that $d = \frac{C}{A} - 1$ (see Section 2.2.3), we obtain that for large sizes the leading order of the mean depth is $d \sim \ln A$, which indicates that what we observe in the simulations is a long transient behavior. This transient behavior leads to the fact that our model fits the data but the asymptotic scaling at the larger sizes will finally be $d \sim \ln A$, as in the ERM.

5.1 Discussion

Different evolutionary models and mechanisms have been proposed in order to explain the branching patterns arising in evolution (Yule,

1924; Aldous, 2001; Blum and François, 2006; Ford, 2006; Stich and Manrubia, 2009; Hernández-García et al., 2010). Here we have introduced a simple model which allows us to estimate the degree of *evolvability*. An increasing number of publications are highlighting the significance of the interplay between evolvability and robustness in evolution (Wagner, 2005; Lenski et al., 2006; Daniels et al., 2008; Wagner, 2008b). By understanding evolvability as the potential of a biological system for future adaptive mutation and evolution (Brookfield, 2009), and robustness as the property of a system to produce relatively invariant output in the presence of a perturbation (Masel and Siegal, 2009), we apply these two concepts to the biological interpretation of the model we propose. In that way, the symmetric diversification event should correspond to the biological context in which the biological system is evolvable, while the asymmetric diversification process should correspond to a biological context where the new biological system, that has just appeared from the diversification process, is robust and unable to diversify forever.

The asymptotic behavior of our model at long sizes recovers the logarithmic behavior of the ERM scaling, so that, as in the models by Stich and Manrubia (2009), the non-ERM behavior occurs as a transient for small tree sizes. Despite this, our aim is to point out the identification of local unbalance in real trees that can be interpreted in terms of the *evolvability* concept. The prevalence of unbalanced branching so found, which is consistent with previous works (Guyer and Slowinski, 1991; Heard, 1992; Guyer and Slowinski, 1993; Mooers et al., 1995; Aldous, 2001; Blum and François, 2006), has been traditionally explained by the presence of variations in the speciation and/or extinction rates all through the Tree of Life (Kirkpatrick and Slatkin, 1993; Mooers and Heard, 1997). Different biological explanations for these variations in the speciation and/or extinction rates have been proposed, such as: refractory period (Chan and Moore, 1999), mass extinctions (Heard and Mooers, 2002), specialization (Kirkpatrick and Slatkin, 1993) or environment effect (Davies

et al., 2005). The consideration of an evolutionary scenario based on the evolvability/robustness interplay has led us to highlight the effect, over the depth scaling, of the presence of asymmetric diversification events, during the evolutionary process, that give rise to a new biological system which is unable to undergo a new diversification event. With the aim of broadening the understanding of the evolutionary processes that lead to such asymmetric diversification events, in Appendix E we expanded the evolvability model including long refractory periods and mass extinction events. This incapability to diversify is a biological phenomenon that takes place at different levels of evolution. So, as we discuss in Appendix E, we can find this incapability at the macroevolutionary level with taxa that require very long refractory periods or with random massive extinctions of taxa, but this incapability can also be reflected at the microevolutionary or at gene level, where the elements unable to diversify are individuals from a population or genetic variants from a cell, embryo or individual.

In summary, the universal scaling properties at gene- and species-level reported in Chapters 3 and 4, strongly suggested the universality of branching rules, and hence of the evolutionary processes that drive biological diversification across the entire history of life, from genes to species. The results presented in the present chapter prove that the topological characterization of the phylogenetic trees can be very helpful in the analysis of the relevance of the robustness of a biological system (species or protein). Thus, the invariance of the scaling properties at both gene- and species-level suggests that the mechanisms leading to the incapability of a biological system to diversify again are present at the rules acting both at genes- and species-level. However, a detailed analysis of branching times will be needed so as to reach a deeper understanding of the diversification processes.

Chapter 6

Depth scaling in taxonomies

Awareness about biodiversity destruction has led to an increasing interest in the study and comprehension of evolutionary processes that give rise to an increase or decrease of biodiversity (Purvis and Hector, 2000; Butlin et al., 2009). In this direction, a large amount of works are integrating taxonomic knowledge on the study of evolutionary patterns that take place in nature (May, 1990; Schwartz and Simberloff, 2001; Samper, 2004; Sahney et al., 2010; Schlick-Steiner et al., 2010). Compared to the widely assumed reliability of those studies that apply the phylogenetic approach to the evolutionary comprehension of biodiversity (Kirkpatrick and Slatkin, 1993; Sanderson, 1996; Mooers and Heard, 1997; Vázquez and Gittleman, 1998; Ricklefs, 2007; Cavender-Bares et al., 2009; Cadotte et al., 2010), the taxonomic application to the comprehension of biodiversity evolution has been followed by several controversies (de Queiroz, 1988, 1997; Benton, 2000; Nixon and Carpenter, 2000; Bryant and Cantino, 2002; Keller et al., 2003; de Queiroz, 2005; Rieppel, 2005, 2006a,b; Hillis, 2007; Ereshefsky, 2007; Yang and Bourne,

2009). One of the major criticisms to taxonomic approaches is derived from the fact that taxonomic classifications are highly biased by the taxonomic criteria used. One of the major consequences derived from this inconvenient is the fact that *taxonomic trees*, the evolutionary trees that reflect the hierarchical nature of taxonomic classifications, are not well-resolved, showing a high presence of polytomies (Yang and Bourne, 2009). This problem is derived from the fact that traditional taxonomy is a rank-based taxonomy, i.e. it is based on predefined taxonomic ranks, so that, regardless of the number of species considered, the number of taxonomic levels is fixed (de Queiroz, 1997). An alternative taxonomic criterion, and the main opponent of the rank-based taxonomic approach is the phylogenetic nomenclature criterion, which consists of a rank-free classification exclusively based on evolutionary criteria (de Queiroz, 1988, 1997, 2005).

Our main goal in this chapter is to consider the branching pattern of the classification trees and to characterize the effects of the increase of polytomies derived from the resolution problems of the rank-based taxonomic criteria. With that aim, we carry out a comparative analysis among rank-based and rank-free taxonomic trees, as well as phylogenetic trees, based on the characterization of the depth scaling behavior, and also on the characterization of the distribution of the polytomies all over the taxonomic trees.[1]

[1]This comparative analysis between taxonomic and phylogenetic trees is based on biological evolutionary trees. In Appendix F we show the same comparative approach for language evolutionary trees.

6.1 Datasets

For this comparative analysis, we used three different kinds of datasets: a rank-based taxonomic tree, a rank-free taxonomic tree and a set of phylogenetic trees.

For rank-based taxonomic tree we used the taxonomic Tree of Life reconstructed from the taxonomic classification database Catalogue of Life: 2010 Annual Checklist (Catalogue_of_Life, 2010). It contains 1,257,735 species grouped among the seven kingdoms of organisms: Animalia, Archaea, Bacteria, Chromista, Fungi, Plantae, Protozoa and Viruses. Catalogue of Life (CoL) is a taxonomic classification based on seven major taxonomic ranks: kingdom, phylum, class, order, family, genus and species. This database allowed us to develop both a global analysis of the whole taxonomic Tree of Life and a comparative analysis among the taxonomic trees of the different kingdoms.

For rank-free taxonomic tree, we used the Tree of Life (ToL) reconstructed by the Tree_of_Life_Web_Project (2010). This database tries to get a complete representation of all the phylogenetic relationships among every species, living or extinct, from the history of life. It constitutes an appropriate case of a rank-free taxonomic tree (Maddison et al., 2007). On December 2nd 2006 we downloaded, from the web page of the Tree_of_Life_Web_Project (2010), the evolutionary tree, with more than 25,500 organisms.

The phylogenetic trees used in our comparative analysis are the set of 5,212 interspecific phylogenetic trees from TreeBASE database used in Chapter 3, downloaded on June 30th 2007.

	TreeBASE	ToL	CoL
τ_A	1.74	1.61	1.89
τ_C	1.53	1.45	1.80

Table 6.1: τ_A and τ_C values of the cumulative complementary distribution functions (CCDFs) for the branch size, $F(A) \sim A^{1-\tau_A}$, and for the cumulative branch size, $F(C) \sim C^{1-\tau_C}$, of TreeBASE, ToL and CoL.

6.2 Results

The comparative analysis of the cumulative complementary distribution functions (CCDFs) for the branch size, $F(A)$, and for the cumulative branch size, $F(C)$, between CoL and Tol taxonomic trees, shows power-law distributions with the forms of $F(A) \sim A^{1-\tau_A}$ and $F(C) \sim C^{1-\tau_C}$ respectively, as described in Chapter 3 for the organism phylogenies (with $\tau_A = 1.74 \pm 0.03$ and $\tau_C = 1.53 \pm 0.02$ values for the interspecific phylogenies). So, on the one hand, CoL displays a wide power-law distribution, $F(A) \sim A^{1-\tau_A}$, characterized by $\tau_A = 1.89 \pm 0.001$, and ToL's branch size distribution shows a power-law tail with the form $F(A) \sim A^{1-\tau_A}$, where $\tau_A = 1.61 \pm 0.004$ (Figure 6.1(a)). On the other hand, in the case of the cumulative branch size distribution, $F(C)$, CoL displays a wide power-law distribution, $F(C) \sim C^{1-\tau_C}$, with value for τ_C of $\tau_C = 1.80 \pm 0.001$, and ToL displays a power-law tail with the form $F(C) \sim C^{1-\tau_C}$, where $\tau_C = 1.45 \pm 0.001$ (Figure 6.1(b)). These exponents differ between ToL, CoL and TreeBASE, with ToL closer to TreeBASE (see Table 6.1).

The depth scaling analysis (Figure 6.2) shows differences between the two different philosophies of taxonomic classifications: rank-based (CoL) and rank-free (ToL) classification. Thus, ToL displays a depth scaling behavior similar to the one exhibited by TreeBASE,

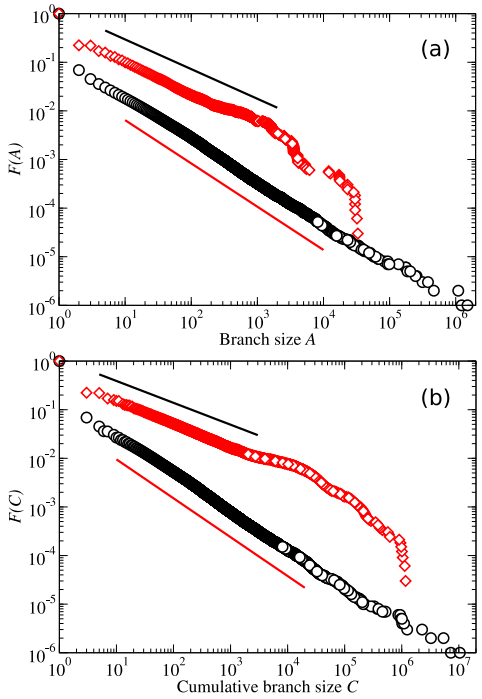


Figure 6.1: Branch size and cumulative branch size distributions of taxonomic trees. (a) Cumulative complementary distribution functions (CCDFs) averaged and logarithmically binned of the branch size for ToL (red empty diamonds) and CoL (black empty circles). Black and red lines correspond to two power laws, $F(A) \sim A^{1-\tau_A}$, with their corresponding exponents given by the best fit to the ToL dataset ($\tau_A = 1.61$) and CoL dataset ($\tau_A = 1.89$), respectively. (b) Cumulative complementary distribution functions (CCDFs) averaged and logarithmically binned of the cumulative branch size for ToL (red empty diamonds) and CoL (black empty circles). Black and red lines correspond to two power laws, $F(C) \sim C^{1-\tau_C}$, with their exponents given by the best fit to the ToL dataset ($\tau_C = 1.45$) and CoL dataset ($\tau_C = 1.80$), respectively.

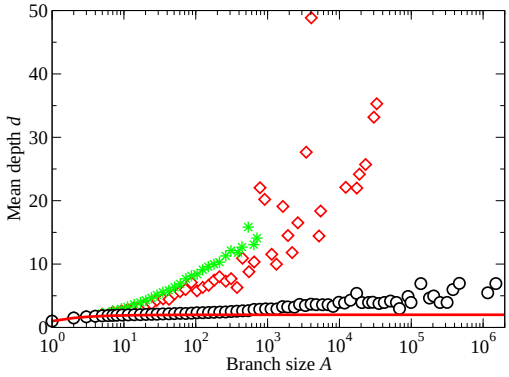


Figure 6.2: Depth scaling of taxonomic and phylogenetic trees. Plot of the logarithmically binned set of the mean depth scaling of CoL (black empty circles), ToL (red empty diamonds) and TreeBASE (green stars). Red line represents the mean depth scaling of the fully polytomic tree.

while CoL displays a depth scaling behavior closer to the depth scaling of the star-like tree, the fully polytomic tree. The resemblance of CoL to the fully polytomic trees is given by the fact that the mean depth, in both cases, is contrained to a fixed number of levels such that, regardless of the size of the tree, the mean depth approaches a constant. For a fully polytomic tree the mean depth is fixed to 2, the mean depth of the rank-based taxonomic tree of CoL is 8 (i.e. the 7 main taxonomic ranks and the root).

In order to characterize if the average behavior described for the whole taxonomic tree of the CoL database is also preserved after discriminating the different kingdoms of organisms, we decided to analyze the scaling of the mean depth as function of the tree size for each of the 8 kingdoms included inside the CoL taxonomic classification (Animal, Archaea, Bacteria, Chromista, Fungi, Plantae, Protozoa and Viruses). The depth scaling analysis for each

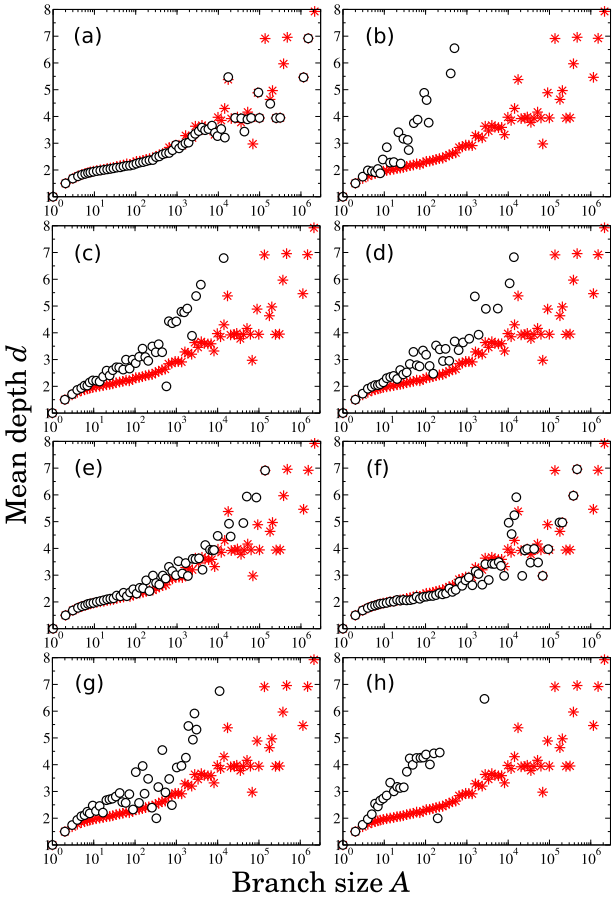


Figure 6.3: Depth scaling of taxonomic kingdoms. Plot of the logarithmically binned set of the average mean depth scaling of the different kingdoms of CoL dataset (black empty circles): Animalia (a), Archaea (b), Bacteria (c), Chromista (d), Fungi (e), Plantae (f), Protozoa (g), Viruses (h). Red stars correspond to the logarithmically binned set of the mean depth scaling of the whole CoL dataset.

taxonomic kingdom shows that several kingdoms display similar depth scaling behavior as the exhibited by the whole CoL dataset, with exception of the kingdoms with low representation (Chromista (11,987 species), Bacteria (11,877 species), Protozoa (9,406 species) and, especially, Viruses (2,039 species) and Archaea (364 species)), which show small deviations from that behavior (Figure 6.3). These kingdoms differ from the average CoL depth scaling behavior and display a much closer behavior to the one displayed by the TreeBASE or ToL datasets.

With the aim of evaluating the presence of polytomies, as well as their distribution all over the taxonomic trees, we analyzed the degree distribution and its dependence with the depth of the taxonomic trees obtained from CoL and ToL databases. The degree depicted the polytomic nature of taxonomic trees, the wider degree distributions of the taxonomies standing out, as compared to the degree distribution of the TreeBASE phylogenies (Figure 6.4(a)). Therefore, in the rank-based taxonomy CoL it is possible to find nodes with degree $k = 2,014$, as well as in the rank-free taxonomy nodes with degree $k = 416$, while for the phylogenetic trees downloaded from TreeBASE, the larger degree that can be observed is $k = 66$. The degree distribution of ToL and TreeBASE can be fitted to a power law distribution with the form $F(k) = k^{\gamma}$, where the value of γ for ToL is $\gamma = -1.76 \pm 0.02$, while for TreeBASE is $\gamma = -2.93 \pm 0.04$. Otherwise, ToL and TreeBASE show an outstanding prevalence of nodes with degree $k = 3$, i.e. binary bifurcations, which reflects their tendency towards a binary topology, as compared to the rank-based taxonomy CoL, which does not show any outstanding prevalence of the binary bifurcations (Figure 6.4(b)).

The analysis of the correlation of the degree with the mean depth implies, once again, a discrepancy between the rank-free and the rank-based taxonomic classification approach. So, while ToL and the TreeBASE phylogenetic trees exhibit no correlation between the degree of the taxon and its depth, with a generic average degree around 3 for all the internal taxa (Figure 6.5(a)), CoL suggested some

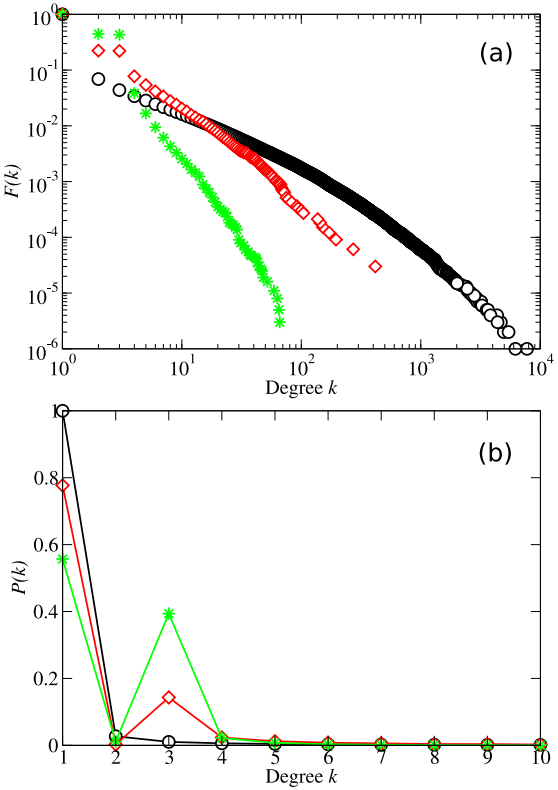


Figure 6.4: Degree distribution of taxonomic and phylogenetic trees. (a) Cumulative complementary degree distribution, $F(k)$, of CoL (black empty circles), ToL (red empty diamonds) and TreeBASE (green stars). (b) Degree distribution, $P(k)$, of CoL (black empty circles), ToL (red empty diamonds) and TreeBASE (green stars) for the range of degree values between 1 and 10.

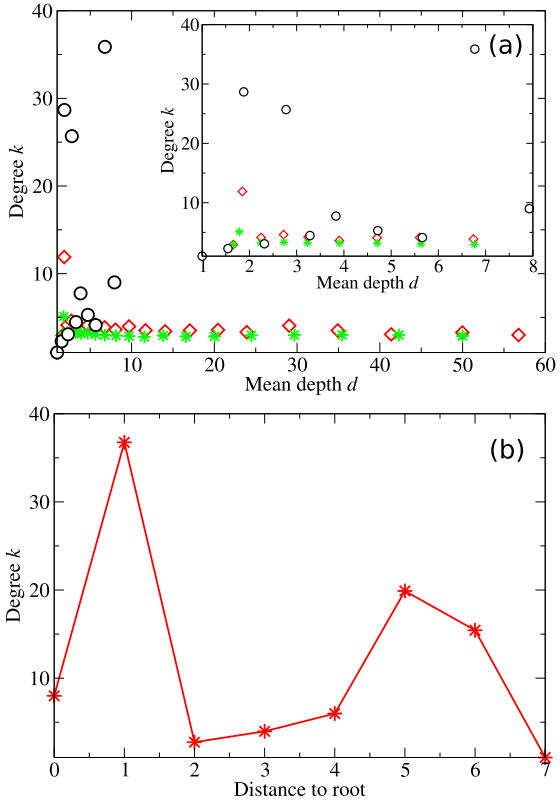


Figure 6.5: Degree-depth correlation of taxonomic and phylogenetic trees. (a) Plot of the logarithmically binned degree-depth correlation of CoL (black empty circles) and ToL (red empty diamonds) and TreeBASE (green stars) datasets. Inset: Enlargement of the range of mean depth values between 1 and 8. (b) Average degree for each of the taxonomic ranks of CoL (red stars): root (0), kingdom (1), phylum (2), class (3), order (4), family (5), genus (6) and species (7).

correlation between the degree of the taxa and their depth. Since in CoL each taxonomic level is at a certain distance from the root, with the aim of expanding the comprehension of the correlation between the degree and the depth of the taxonomic tree, in Figure 6.5(b) we show the correlation of the degree with the taxonomic rank. The plot shows significantly higher degrees at those taxonomic ranks with distances to the root of 1, 5 and 6, corresponding to kingdom, family and genus, respectively.

With the aim of determining which is the specific contribution of each kingdom to the degree-taxonomic rank correlation depicted for the whole CoL taxonomic tree, we discriminated among the different kingdoms of organisms. In this way, as we can see in Figure 6.6, the degree-depth correlation analysis for the different kingdoms shows that the high polytomic nature of family and genus is especially supplied by those kingdoms with large representation, i.e. Animalia (1,391,352 species), Plants (448,695 species) and Fungi (125,205 species). Otherwise, the main contributors to the extremely high polytomic nature of the taxonomic rank of kingdom are the kingdoms of Protozoa and Viruses, together with the kingdoms of Animalia, Plantae and Bacteria. Another remarkable finding is the low contribution of Archaea kingdom to the polytomic topology of CoL taxonomic tree. This can be related to its extraordinary low representation, consisting only of 364 species.

6.3 Discussion

Ever since the characterization of the uneven distribution of species inside genera by J. C. Willis in 1922 (Willis, 1922), a large amount of studies have focused on the characterization of the distribution of subtaxa inside taxa all over the Tree of Life, such as Corbet (1942); Anderson (1974); Burlando (1990, 1993). In the last decades diverse

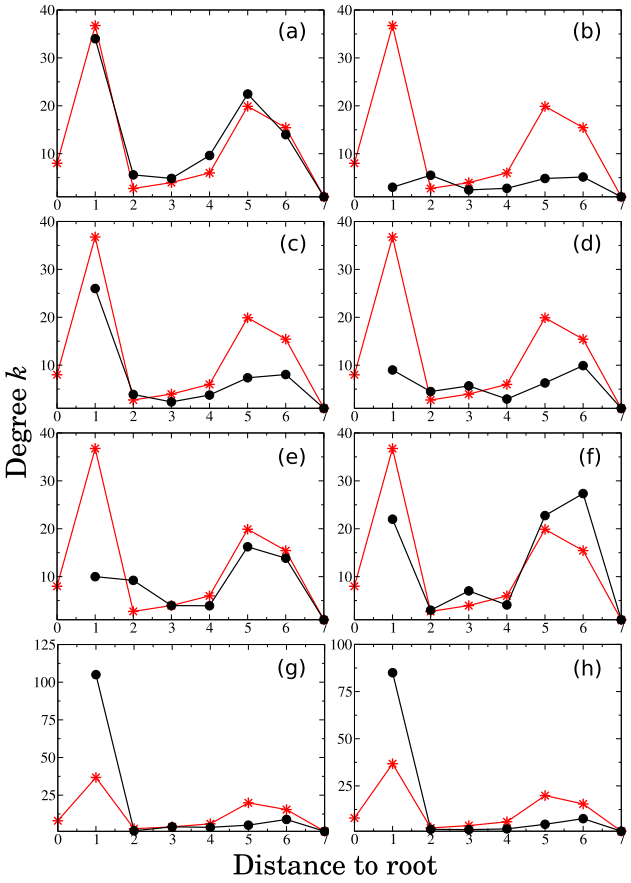


Figure 6.6: Degree-depth correlation of taxonomic kingdoms. Plot of the logarithmically binned degree-depth correlation of the different kingdoms of CoL dataset (black solid circles): Animalia (a), Archaea (b), Bacteria (c), Chromista (d), Fungi (e), Plantae (f), Protozoa (g), Viruses (h). Red stars correspond to the logarithmically binned set of the degree-depth correlation of the whole CoL dataset.

studies based on the modern network theory have been carried out for the analysis and comprehension of the biodiversity patterns printed in the taxonomic classifications. The major studies can be classified into two main approaches. On the one hand, several studies have been published with a focus on the characterization of the taxon abundance, with the aim of understanding the distribution of subtaxa inside their respective taxa, e.g. species inside genera, or genera inside families (Chu and Adami, 1999; Reed and Hughes, 2002; Caldarelli et al., 2004; Reed and Hughes, 2007). On the other hand, some analyses have focused on the characterization of the distribution of polytomies, through the analysis of the degree distribution of the taxa (Cartozo et al., 2008). Taking into account the great controversy between rank-free and rank-based taxonomic classification criteria (de Queiroz, 1988, 1997; Benton, 2000; Nixon and Carpenter, 2000; Bryant and Cantino, 2002; Keller et al., 2003; de Queiroz, 2005; Rieppel, 2005, 2006a,b; Hillis, 2007; Ereshefsky, 2007; Yang and Bourne, 2009), in this chapter we decided to characterize the effect of the different taxonomic classification criteria over the branching properties of the evolutionary trees.

Our characterization of the effect of the criteria of taxonomic classification over the topology of the evolutionary trees focused on the characterization of the depth scaling behavior and in the distribution of the polytomies all over the taxonomic trees. In general, for all the analysis of both taxonomic datasets, ToL and CoL, ToL was the one which showed closer behavior to phylogenetic trees. The characterization of the distribution of subtaxa inside taxa, through the analysis of the branch size and cumulative branch size distribution, displayed a power-law distribution for both taxonomic datasets (see Figure 6.1), in line with the large amount of analyses published in this direction (Willis, 1922; Corbet, 1942; Anderson, 1974; Burlando, 1990, 1993; Chu and Adami, 1999; Reed and Hughes, 2002; Caldarelli et al., 2004; Reed and Hughes, 2007). However, ToL's distributions showed closer exponents to the ones reported in Chapter 3 for the branch size and cumulative branch size distribution analysis of or-

ganism phylogenies than CoL's distributions did. In addition, the closeness of the ToL dataset to phylogenetic trees became also evident in the mean depth scaling analysis (see Figure 6.2), as well as in the degree distribution (see Figure 6.4).

Some years ago, Cartozo et al. (2008) published a paper based on the characterization of the degree distribution of the taxonomic tree of the phylum of vascular plants. The taxonomic classification was not constrained by the taxonomic ranks considered by CoL, including intermediate ranks, so the taxonomic ranks that they considered in this work were: species, genus, family, order, sub-class, class, subphylum, phylum and subkingdom. They reported that the degree follows a power-law distribution in taxonomies. This result does not agree well with the distribution displayed by CoL taxonomy, but it matches up with the distribution described for the rank-free taxonomy ToL (see Figure 6.4). Since the taxonomic classification considered by Cartozo et al. (2008) took into account intermediate taxonomic ranks, by increasing the resolution of the taxonomic classification (i.e. including intermediate ranks or the consideration of rank-free criteria), the power-law distribution described by Cartozo's dataset and by ToL and TreeBASE datasets could be explained by the softening of the constraints imposed by the main seven taxonomic ranks of the rank-based taxonomic classification (kingdom, phylum, class, order, family, genus, specie).

An interesting point in the comprehension of the distribution of polytomies all over the taxonomic trees is the analysis of the degree of correlation of the distribution of polytomies with the depth of the taxa. In that sense, our degree-depth correlation analysis showed no correlation for ToL, as was also the case for TreeBASE, but CoL displayed a correlation between the number of polytomies and the taxonomic level, and it exhibited especially larger average degree values at the taxonomic ranks corresponding to genus, family and kingdom (see Figure 6.5). This result implies that the higher presence of polytomies and, therefore, a higher presence of low resolution

problems at the taxonomic classification is found at those taxonomic levels: genus, family and kingdom.

The specific analysis of each of the kingdoms of organisms demonstrates that the just described perturbing effect of the rank-based criteria over the tree topology of the whole taxonomic classification of CoL is not reflected at the different kingdoms with the same intensity. As a consequence, while the largely represented kingdoms (Animalia, Plantae and Fungi) displayed a depth scaling behavior closer to the depth scaling of the fully polytomic tree, the less represented kingdoms (Archaea and Viruses) drew away from the average behavior to a depth scaling behavior closer to the one described by the rank-free taxonomy of ToL. These discrepancies show the implication of the rank-based criteria over the high presence of polytomies in taxonomies like CoL. Since CoL is restricted to 7 taxonomic ranks, regardless of the number of species inside a kingdom, the more species a kingdom has, the more polytomic behavior its taxonomic tree will depict. In addition, the analysis of the correlation of the degree with the taxonomic rank for the different kingdoms shows that, while Archaea, due to its low representation, showed a low contribution for the polytomic topology of CoL, those kingdoms with a larger representation (Animalia, Plants and Viruses) bear the responsibility for the polytomic nature of family and genus ranks, as well as they are the main contributors to the extremely high polytomic nature of the taxonomic rank of kingdom. These kingdoms are, specifically, the kingdoms of Protozoa and Viruses, along with the kingdoms of Animalia, Plantae and Bacteria.

In summary, the main aim of this work has been the characterization of the effect of the different taxonomic criteria over the topology of the evolutionary trees. The results presented in this chapter show that rank-based criteria distort considerably the branching pattern of the evolutionary trees, as compared with phylogenies, while the rank-free taxonomic criteria are more respectful with the pattern. The perturbation of the topological properties of the rank-based taxonomic trees is caused by the fact that the depth of such trees is

constrained by a fixed number of taxonomic ranks. This constraint is what causes the increase in the number of polytomies. Based on the results obtained here, we can conclude that, from the evolutionary point of view, rank-free classification criteria are more respectful with the evolutionary patterns reflected through the branching patterns of the evolutionary tree than the rank-based taxonomic criteria and therefore, the rank-free approach constitutes a more realistic taxonomic classification than the rank-based one.

Chapter 7

Branch length scaling

As discussed in Chapter 2, phylogenetic trees provide us with two major sources of information (Moore, 2007): branching and temporal information. Traditionally, most of the topological characterization studies have focused on the branching pattern analysis, while the number of works that take into account the information stored in the branch length of the phylograms is much shorter (Savage, 1983; Kirkpatrick and Slatkin, 1993; Mooers and Heard, 1997). The sensitivity of the branch length to the reconstruction and timing methods has made that most of the studies based on branch length data have focused on the identification of biases derived from the reconstruction process of the phylogenetic trees (see Section 2.1). But the improvement in the dating methods in the last decade has given rise to the development of databases that provide information, with considerable precision, about the branch length of the phylogenetic trees. An example of those databases is the TimeTree project, which is in charge of the time calibration of the Tree of Life and which provides the branch length information (Hedges et al., 2006; Timetree, 2010). With the aim of contributing to the comprehension of the evolutionary processes that drive branch length patterns, we

analyzed the TimeTree of Life database to characterize the branch length pattern along the Tree of Life.

7.1 Datasets

On September 13th 2010 we digitalized the circular timetree displayed as wall poster at Hedges et al. (2006) and Timetree (2010). Nowadays, Timetree of Life database constitutes the more accurate calibration to absolute time of the diversification events depicted all over the Tree of Life, providing information about the timescales of the evolutionary processes among all the life forms included in the Tree of Life. Since it is a work in progress, the available data are increasing in time. The version downloaded on September 13th 2010 included all three superkingdoms (Eubacteria, Archaea and Eukarya) and 1,610 families. Timetree of Life resolution stops at family-level taxa, therefore, the 1,610 families constitute the leaves of the downloaded Timetree.

7.2 Results

The branch length frequency distribution is quantitatively characterized by calculating the function $F(L)$, which is the complementary cumulative distribution function (CCDF) of the branch lengths (in million years (My)), L, in the TimeTree. The branch length distribution of the TimeTree displays an exponential decay with two different decay rates (Figure 7.1(a)), with an inflection point around 200 and 300 million years length (inset of Figure 7.1(a)). The detailed analysis of the branch length distribution for the main groups

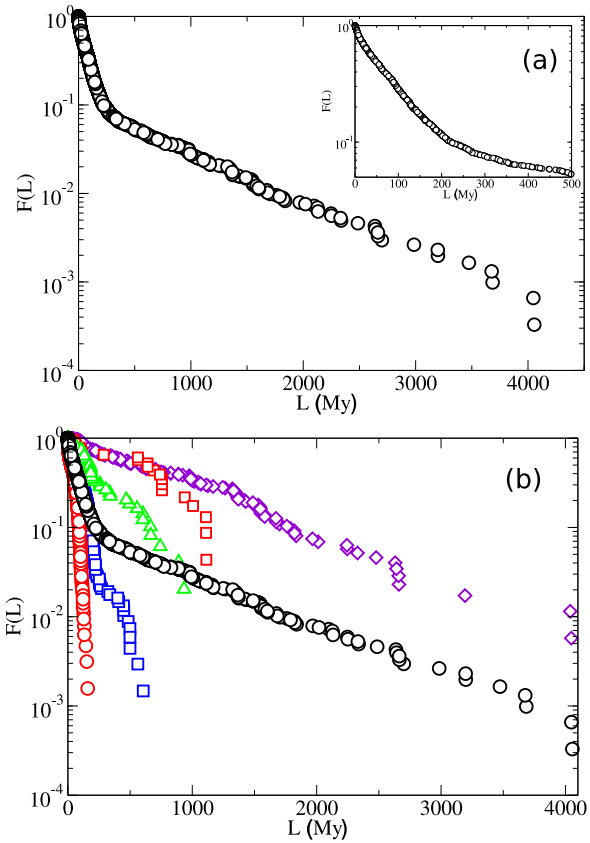


Figure 7.1: Branch length distribution. (a) Cumulative complementary distribution function (CCDF) of the branch lengths of Timetree (black empty circles). Inset: Zoom in of the cumulative distribution of branch lengths shorter than 500 million years old. (b) CCDF of the branch lengths for eubacteria (violet empty diamonds), fungi (red empty squares), protists (green empty triangles), arthopods (blue empty squares), flowering plants (red empty circles) groups of organisms in Timetree. Black empty circles correspond to the CCDF of the branch lengths of the whole Timetree.

of organisms in the Tree of Life (amphibians, arthropods, birds, eubacteria, ferns, fishes, flowering plants, fungi, liverworts, mammals, mollusks, mosses and protists) shows exponential distributions with a single characteristic time for each of those groups (Figure 7.1(b)). Therefore, the described double exponential distribution is explained by the different decay rates of the different groups of organisms. So, while most of the groups show a fast exponential decay (mosses, liverworts, flowering plants, arthropods, fish, amphibians, reptiles, birds and mammals), some others show significantly slower exponential decays, such is the case of eubacteria, archaea, protists and fungi.

In order to see how the branch lengths are distributed all over the Timetree of Life, we plot the length of each of the branches that we find at each time (Figure 7.2(a)). This plot shows a positive correlation of branch length with age. So, ancient branches are more likely to be long than young branches. This behavior contrasts with the one characteristic for random trees, in which the length of their branches showed no correlation with their age (inset of Figure 7.2(a)). The positive correlation is softer if we include the branches that connect leaves in the analysis since there are some of those branches that bifurcated very soon, and which have been present most of the time in the tree and, therefore, they tune down the positive correlation with age. In Figure 7.2(b), we can see that while most of the leaves are connected by short branches, numerous branches are very long, responsible for tuning down the time correlation.

7.3 Discussion

The availability of branch length information in the last decades has led to an increase in the number of studies focused on the analysis of how the branch lengths are distributed all over the phylogenetic

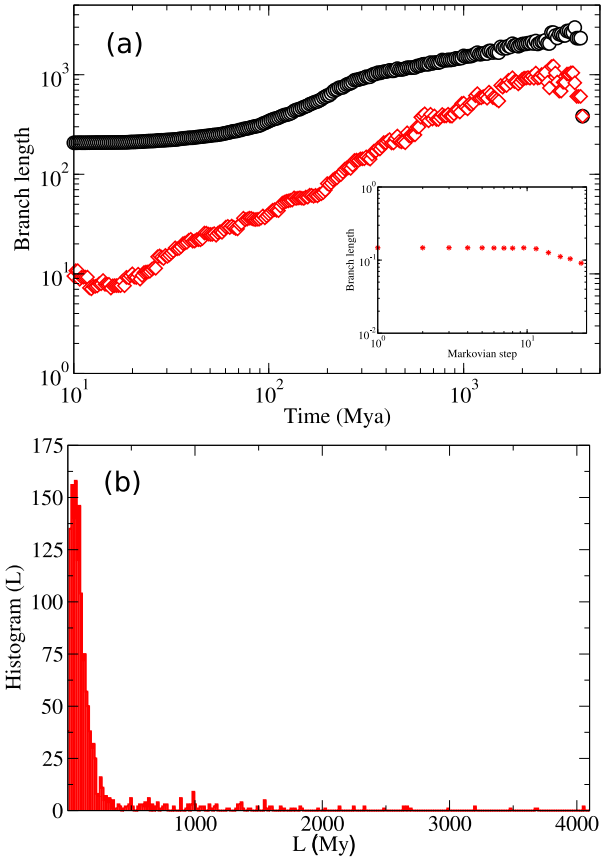


Figure 7.2: Branch length-time correlations. (a) Plot of the logarithmically binned set of branch lengths for each time (in million years ago (Mya)) for the whole Timetree (black empty circles) and for the Timetree without leaves (red empty diamonds). Inset: Plot of the logarithmically binned set of the branch lengths for each Markovian step for generated random trees. (b) Histogram of the lengths of the branches that connect the leaves of Timetree.

trees (Fiala and Sokal, 1985; Rohlf et al., 1990; Zink and Slowinski, 1995; Salisbury, 1999; Pybus and Harvey, 2000; Nee, 2001; Popovic, 2004; Qiao et al., 2006; François and Mioland, 2007; Paradis, 2008; Venditti et al., 2010). While the predominance of the analysis of the branching patterns has led to a large amount of information about the biological mechanisms that contribute to the branching patterns of the phylogenetic trees, the description of the biological mechanisms that take part in the branch length patterns of the Tree of Life is still just incipient (Mooers and Heard, 1997). With the aim of broadening the knowledge of the phylogenetic branch length patterns, we characterized the branch length distribution all over the Timetree of Life.

The frequency distribution of the branch lengths in the Timetree of Life displayed an exponential decay with two different decay rates, a lower rate for branch lengths longer than around 200-300 My, and a higher rate for branch lengths shorter than around 300-200 My. These different decay rates are due to the different exponential behaviors of the different groups of organisms: those groups of organisms that diversified in the last 500 My (mosses, liverworts, flowering plants, arthropods, fishes, amphibians, reptiles, birds and mammals) show higher exponential decays, while those groups of organisms that diversified before 500 Mya (eubacteria, archaea, protists and fungi) exhibit lower exponential decays. The exponential frequency distribution observed here confirms the results published by Venditti et al. (2010), who described an exponential frequency distribution of the branch lengths for 101 phylogenies from a narrow taxonomic range of species. Besides, the distribution of the branch length showed a non-random stemmy behavior with a positive correlation with age (see Figure 7.2(a)), short branches being more likely to be found close to the tips, while long branches are close to the root.

From the biological point of view, a possible explanation for the high frequency of the branch lengths shorter than 200-300 My and for the stemminess of Timetree could be that both are consequences

of faster diversification rates of the recent taxa. In fact, based on the stemmy nature of the Timetree, most of the branches shorter than 200-300 My are close to the tips. And, taking into account that the land colonization in the late Paleozoic Era (359-251 Mya) comprised an extraordinary diversification rate increase, we could relate the change in the exponential frequency distribution with such an increase in the diversification rate (Hedges and Kumar, 2009). However, before considering these sorts of biological explanations, we have to take into account some methodological caveats claimed by the authors of this first version of Timetree database (Hedges and Kumar, 2009). First of all, the analyzed Timetree version includes only living organisms sampled by molecular methods, which involves the exclusion of extinct taxa. And, since older taxa have more time to go extinct than recent taxa, the absence of extinct taxa leads to stemmier evolutionary trees (Nee et al., 1994a). Another problem, derived from this first caveat, is that, due to a limited availability of molecular data for certain groups of organisms, the different taxa are unequally covered. Third, taxonomic criteria are very arbitrary with evolutionary time and, therefore, there is no guarantee that a group of organisms that corresponds to a certain taxonomic rank can be compared to a different group of organisms of the same taxonomic rank (Barraclough and Nee, 2001; Avise and Mitchell, 2007; Hedges and Kumar, 2009). An example of a consequence derived from these drawbacks could be the high frequency of long branches in the old groups of organisms (eubacteria, archaea, protists and fungi), which could be explained as the result of wrong taxonomic criteria and unequal coverage of those taxa, rather than as the result of low diversification rates.

Summing up, as a first step towards the characterization of the branch length pattern along the Tree of Life, we have confirmed the exponential frequency distribution of the branch lengths in the Tree of Life, as well as the stemminess pattern of the Timetree depicted in the positive age correlation of the branch length distribution. Nonetheless, we should wait for improved versions of the Timetree

database for extracting round biological conclusions from this kind of analysis.

Chapter 8

Conclusions

Within the context of complex network theory, the main purpose of this work has been to provide new tools for the topological characterization of evolutionary trees, as well as to contribute to enlarging the knowledge of the evolutionary patterns depicted in them. Thereby, we have proposed an approach based on the depth scaling analysis of phylogenetic trees in order to carry out comparative studies between micro- and macroevolutionary phylogenies, gene and species evolutionary trees, as well as a comparative study of the effects of the rank-based and rank-free taxonomic criteria over the topology of evolutionary trees. Besides, we have examined the information stored in the branch length of the phylograms, in order to characterize the branch length distribution along the Tree of Life.

In the first place, we applied, in Chapter 3, the depth scaling approach to the characterization of intra- and interspecific phylogenies across kingdoms, reproductive strategies and environments. This analysis showed similar branching and scaling patterns, both in intraspecific and interspecific phylogenies, suggesting the conservation of branching rules and, hence, of evolutionary processes that drive biological diversification across the entire history of life.

The deviation from the null ERM model of the phylogenies observed suggests the operation of a mechanism generating a correlated branching, where some memory of past evolutionary events is maintained along each branch.

With the aim of checking if the universality found in the scaling properties of organism phylogenies is also extrapolable to the gene-level, in Chapter 4 we characterized the depth scaling of protein families. The results presented there showed that the branching and scaling patterns in protein families do not differ significantly from the branching patterns observed in organism phylogenies, suggesting that the branching rules and evolutionary processes that drive biological diversification are conserved both in species- and in gene-level.

As a proposal of an evolutionary process that can be conserved in gene- and in species-level, in Chapter 5 we introduced an evolutionary model based on the evolvability/robustness interplay of genes or organisms. This model showed the ERM scaling as asymptotic behavior, but there was a distinct non-logarithmic behavior for finite sizes. It reproduced the branching and scaling properties displayed by the phylogenies analyzed in Chapters 3 and 4, showing a prevalence of asymmetric diversification events, i.e. events in which one of the two daughter species or genes is unable to diversify for a very long time or at all.

In Chapter 6 we characterized the effect of the different taxonomic classification criteria over the branching properties of the evolutionary trees. The analysis showed a very distorting effect of the rank-based taxonomic criteria over the biodiversity distribution through the evolutionary trees, while the rank-free criteria proved to be more respectful with it. This perturbation effect was especially reflected in a high presence of polytomies in the taxonomic ranks of genus, family and kingdom, as well as on the depth scaling of the kingdoms of organisms with larger representation such as Animalia, Plantae and Fungi. We note that the distinct behavior of taxonomies and

phylogenies revealed by our topological methods confirms that the universality of results reported in Chapters 3 and 4 is not a consequence of the analysis methodology used, but it is a property of the biological datasets.

Finally, in Chapter 7 we took a first step toward the characterization of the branch length pattern all over the Tree of Life, and we reported the exponential frequency distribution of the branch length in the Timetree of Life, as well as the stemminess pattern of this Tree of Life. This branch length distribution can be associated with faster diversification rates of recent taxa, but some caveats derived from the reconstruction of the first version of the Timetree database oblige us to be cautious and not to draw any definite conclusion before analyzing improved versions of the Timetree database which still have to be published.

Part I

Appendices

Appendix A

Intra- and interspecific datasets

The interspecific dataset analyzed in Chapter 3 consists of 5,212 phylogenetic trees downloaded from TreeBASE (2010). Given that a database similar to TreeBASE does not exist for intraspecific phylogenies, we constructed our intraspecific dataset by manually compiling 67 phylogenetic trees from several published references (see Table A.1). The difference in size between the two datasets calls for some additional checking on the appropriateness of a comparison between them. As a way to close the gap between the two datasets we compiled a third set of trees consisting of phylogenies of interspecific character, like the data in TreeBASE, but these were manually extracted from published references (see Table A.2) following the same criteria as the intraspecific set analyzed in Chapter 3, and with the same size, 67 trees. Our selection criteria insure that our tree datasets contained organisms from terrestrial, marine and fresh water environments, from all the main climatic regions, from

Animalia	Devitt (2006); Driscoll and Hardy (2005); Heilveil and Berlocher (2006); Jensen et al. (2002); Kawamoto et al. (2007); Lefébure et al. (2006); Marmi et al. (2006); Martínez-Solano et al. (2006); Miller et al. (2006); Ozeki et al. (2007); Roberts (2006); Rowe et al. (2006); Ruzzante et al. (2006); Ursenbacher et al. (2006); Verovnik et al. (2004); Zink et al. (2006)
Bacteria	Ehling-Schulz et al. (2005); Hahn et al. (2005); Hommais et al. (2005); Humbert et al. (2005); Ko et al. (2003); Lin et al. (2003); Sogstad et al. (2006); Vancanneyt et al. (2006); Ward et al. (2004); Zorrilla et al. (2003)
Fungi	Choi et al. (2006); Marimon et al. (2006); Perneel et al. (2006); Scott and Chakraborty (2006)
Plantae	Albach et al. (2006); de Casas et al. (2006); Huang et al. (2001); van Ee et al. (2006)
Protozoa	Beszteri et al. (2005); Cupolillo et al. (2003); Monis et al. (2003); Thangadurai et al. (2006); Whipps and Kent (2006); Zhang et al. (2006b)
Viruses	Gottschling et al. (2007); Michitaka et al. (2006); Perk et al. (2006); Zhang et al. (2006a)

Table A.1: Intraspecific phylogenies datasets.

all kingdoms (see Table 3.1), and they were reconstructed with the main phylogenetic tree estimation methods.[1]

[1]The 134 phylogenetic trees that constitute our manually compiled dataset of intraspecific and interspecific phylogenies are available at: `http://ifisc.uib-csic.es/~alejandro/phylotreedata/`.

Animalia	Benz et al. (2006); Dohrmann et al. (2006); Duda and Kohn (2005); Fuchs et al. (2007); Fulton and Strobeck (2006); Gaubert and Cordeiro-Estrela (2006); Lavoué et al. (2007); Le et al. (2006); Mallatt and Giribet (2006); Mann et al. (2006); Maraun et al. (2004); Moyle and Marks (2006); Ohlson et al. (2007); Robalo et al. (2007); Sagegami-Oba et al. (2007); Sjölin et al. (2005); Sullivan et al. (2006); Zanatta and Murphy (2006); Zuccon et al. (2006)
Archaea	Dighe et al. (2004); Garcia et al. (2000); Wright (2006)
Bacteria	Brindefalk et al. (2007); Huang et al. (2005); Zhang et al. (2001)
Fungi	Fitzpatrick et al. (2006); García et al. (2006); Stchigel et al. (2006); Wang et al. (2006)
Plantae	Andreasen and Bremer (2000); Ellison et al. (2006); Endress and Doyle (2007); Gamage et al. (2006); Hahn (2002); Hyvönen et al. (2004)
Protozoa	Gast (2006); Li et al. (2006); Moreira et al. (2007); Saldarriaga et al. (2003); Sørensen and Giribet (2006)
Viruses	Habayeb et al. (2006)

Table A.2: Interspecific phylogenies datasets.

Appendix B

Outgroup effect over the allometric scaling of phylogenies

The results presented in Chapter 3 were published in 2008 highlighting the non-random universal patterns of phylogenetic differentiation, and suggesting the relevance of similar evolutionary forces that drive diversification across the broad range of scales, from microevolutionary to macroevolutionary processes, shaping diversity of life on Earth (Herrada et al., 2008). Almost a year after the publication of this work, a paper was published criticizing the universality of our results and arguing that the alleged universal behavior could be caused by the effect of the artifacts introduced by the outgroups over the allometric scaling of the phylogenetic trees (Altaba, 2009).

Based on that criticism, we have checked the effect of the outgroups over the allometric scaling of the phylogenetic trees. With that purpose, we collected 9 phylogenies from published research papers (Spinks and Shaffer (2005); Rowe et al. (2006); Fuchs et al. (2008,

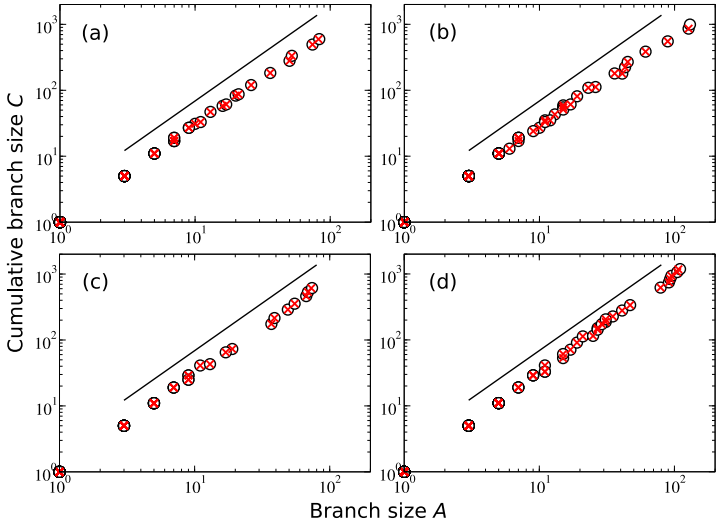


Figure B.1: Outgroup effect over the allometric scaling. Comparative allometric scaling plot between complete phylogenies (black empty circles) and their corresponding ingroup taxa (red crosses) for the phylogenies published in: Fuchs et al. (2008) (a), Fuchs et al. (2009) (b), Yi et al. (2009) (c) and Wahrmund et al. (2010) (d). Line corresponds to the power law $C \sim A^{\eta}$, with the exponent value of $\eta = 1.44$.

2009); Pyron and Burbrink (2009); Yi et al. (2009); Wahrmund et al. (2010); Wilson et al. (2009)) and we contrasted them with the results published by Altaba (2009).

The comparison between the allometric scaling of each phylogenetic tree with the allometric scaling of its corresponding phylogenetic tree without outgroups displays no differences. It also shows the same scaling behavior as the one described in Chapter 3. In Figure B.1 we show the allometric scaling for 4 of the phylogenies analyzed. Complementing this result, the plot of the allometric scal-

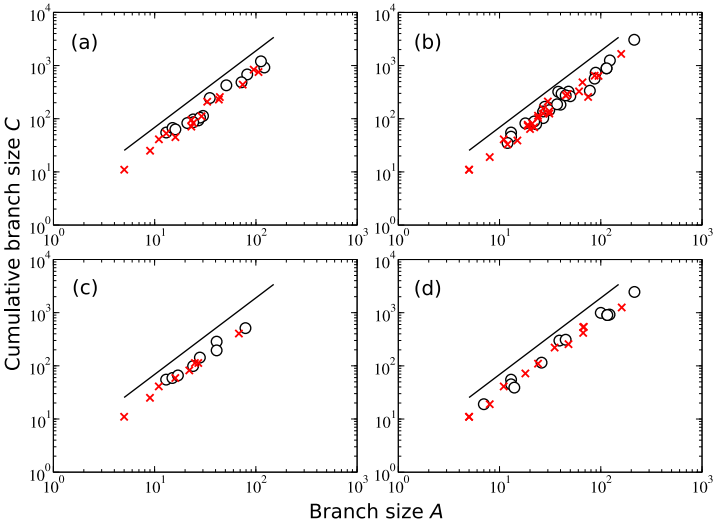


Figure B.2: Outgroup effect over the allometric scaling based on the data published by Altaba (2009). Comparative allometric scaling plot for the values of the branch size, A, and the cumulative branch size, C, corresponding to the root of the complete phylogenies (black empty circles) and to the root of their corresponding ingroup taxa (red crosses), published in Altaba (2009). (a) Phylogenies reconstructed by maximum likelihood. (b) Phylogenies reconstructed by maximum parsimony. (c) Phylogenies reconstructed by neighbor joining. (d) Phylogenies reconstructed by bayesian inference. Line corresponds to the power law $C \sim A^{\eta}$, with the exponent value of $\eta = 1.44$.

ing relationship between the cumulative branch size, C, and branch size, A, values for the roots of the complete phylogenies and ingroup taxa analyzed in Altaba (2009) shows also the same scaling behavior as the one described in Chapter 3 (Figure B.2).

The main function of the outgroup is rooting the ingroup taxa, which means that the difference between both sets of data (whole phylogenies and its corresponding ingroups) is due to a few points concerning the internal nodes that connect the outgroup(s) with the root. The results presented here demonstrate that the absence of those nodes does not affect the allometric scaling of the phylogenies, denying the presence of artifacts introduced by the outgroups over the allometric scaling of the phylogenetic trees, as was claimed by Altaba (2009), and reaffirming the universality in the allometric scaling of the phylogenetic trees described in Chapter 3.

Appendix C

Orthologs-paralogs correlation

As we described in Section 1.1.2, the two main processes responsible for protein family diversification are speciation (orthology) and gene duplication (paralogy) (Koonin, 2005). In order to broaden the understanding of protein family evolution, in the present appendix we compute which of those diversification processes contributes more to protein family diversification, through the measure of the number of species per protein family as function of the protein family size. For a protein family that diversifies only by speciation events (fully-orthologous protein family) the number of species inside the protein family increases directly proportional to the size of the family, while for a protein family that diversifies exclusively by gene duplication events (fully-paralogous protein family) the number of species inside the family remains constant to 1, since all the diversification events due to gene duplication take place in the same species. Taking this into account, we plot how the number of species inside each family changes with the size of the protein family for the "real" protein families.

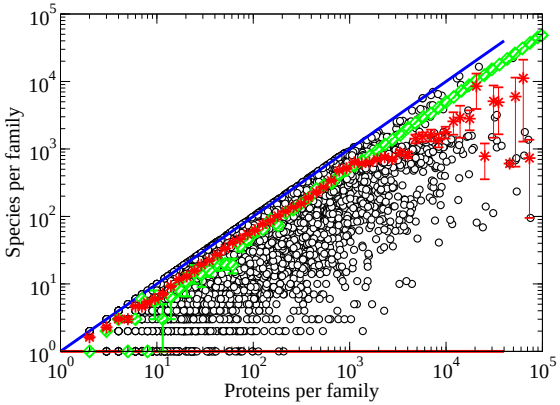


Figure C.1: Number of species per protein family size. Plot of the number of species as function of the protein family size for each the protein families of Pfam database (black empty dots). Red stars and green patterned diamonds represent the logarithmically binned set of values of the number species as function of the protein family size for the Pfam protein families and for the random model (see text), respectively. Blue and red lines represents the extreme behaviors corresponding, respectively, to the fully-orthologous ($y = x$) and fully-paralogous ($y = 1$) protein families.

The protein families analyzed in this study were the protein families from the database Pfam (Pfam, 2010). On November 7th 2008, we downloaded the 8,192 protein families that were collected at this time in Pfam database. The size of the protein families ranges from 3 to more than 2,000 members.

In Figure C.1 we plot the number of species as function of the protein family size. While the number of species per protein family size for all the Pfam families were enclosed between the two extreme behaviors (those corresponding to fully-orthologous and fully-paralogous

protein families), the mean values show an intermediate behavior and they are close to the average behavior of a random uniform model. In this random model, for each protein family size, a certain value for the number of species was randomly chosen from an uniform distribution inside the range that corresponds to the range enclosed between the two extreme behaviors, $y = x$ and $y = 1$ that is $[1, T]$, where T corresponds to the protein family size.

The result presented here does not give any conclusive idea about the prevalence of the gene duplication over the speciation, or vice versa, in the protein family diversification. The coincidence of Pfam protein families with the random model may suggest that diversification of protein families could be the result of a uniform combination of both evolutionary mechanisms, without forgetting other evolutionary mechanisms that take part of protein diversification processes (see Section 1.1.2), such as: horizontal gene transfer (xenology), barrier to sex chromosome recombination (gametology), whole-genome duplication (ohnology) or hybridization of two species (synology).

Appendix D

Activity model

The activity model constitutes an example that tree shapes distinct from the ERM model may also result from a memory in terms of internal states of the nodes. This model is conceptually similar to the class of models suggested by Pinelis (2003). However, the present model makes a distinction only between active and inactive nodes and has a single parameter controlling the spread of activity.[1]

Starting from a single node (the root), a binary tree is generated as follows. At each step, a leaf i of the tree is chosen and branched into two new leaves. Each of the two new leaves, independently of the other, is set active with probability p or inactive with probability $1 - p$. The branching leaf, i, is chosen at random from the set of active leaves if this set is non-empty. Otherwise, i is chosen at random from the set of all leaves. Figure D.1 shows that for $p = 1/2$ the model generates trees with mean depth growing as the square root of tree size (note the log-log scale). Figure D.2 displays a small-size example of such trees and examples from other models, such as the ERM model (Figure D.2 a)) and the alpha model (Figure D.2

[1] The work described in this appendix has been published in (Hernández-García et al., 2010).

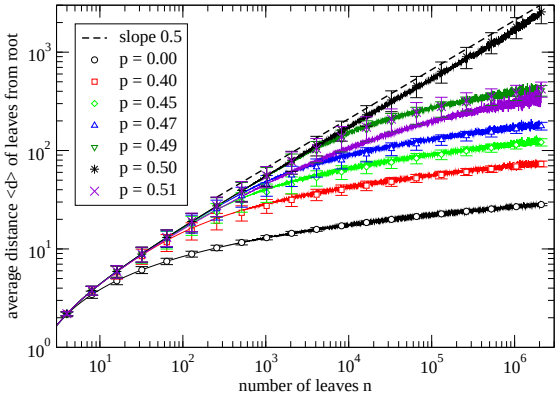


Figure D.1: Average depth versus size for the activity model for various values of the activation probability p. Data points displayed by symbols give the average distance of leaves with respect to the root. Error bars give the standard deviation taken over different realizations (1000 trees per data point). Data in the rugged curves are for all subtrees of trees with size $2^{21} = 2097152$. The dashed line represents a power law scaling with exponent 1/2, corresponding to the scaling of the $p = 0.5$ curve, as discussed in the text.

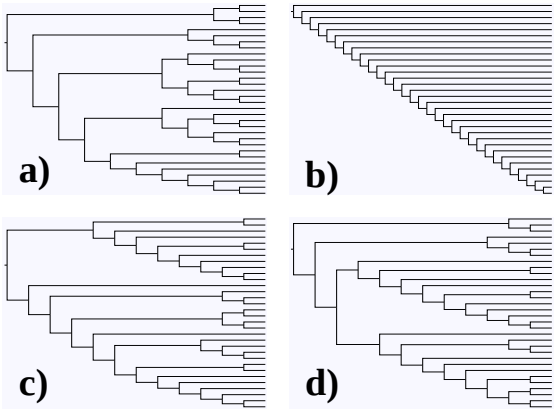


Figure D.2: Examples of trees with 32 leaves, generated from several models. a) Tree generated with the ERM model. b) The completely unbalanced tree. c) A tree generated with the alpha model (described in Subsection 2.3.2) for $\alpha = 0.5$. d) A tree generated with the activity model for $p = 0.5$. The trees in c) and d) display an imbalance intermediate between a) and b).

c)). For values of p below or above $1/2$, $\overline{N}$, the average distance from the leaves to the root (defined in Section 2.2.1), seems to increase logarithmically with n.

Here we give a simplified argument to understand the observed exponent $1/2$ of the distance scaling with system size in the case $p = 1/2$. At the time the growing tree has n leaves in total, let $D_a(n)$ be the expected sum of distances of active leaves from the root, and $D_b(n)$ the analogous quantity for the inactive leaves. When a randomly chosen active leaf –at distance d_a from root– branches, the expected increase of $D_a(n)$ is

$$\begin{aligned} \Delta D_a(n) &\equiv D_a(n+1) - D_a(n) = \\ p^2(d_a+2) &+ 2p(1-p)\cdot 1 + (1-p)^2(-d_a) \\ &= (2p-1)d_a + 2p\,. \end{aligned} \tag{D.1}$$

Here the three terms of the second line are for the activation of two, one and zero of the new leaves, respectively. This expression is appropriate as far as the number of active nodes is not zero. Simultaneously, the expected change in $D_b(n)$ during the same event is

$$\begin{aligned} \Delta D_b(n) &= \\ p^2\cdot 0 &+ 2p(1-p)(d_a+1) + (1-p)^2 2(d_a+1) \\ &= 2(1-p)(d_a+1)\,. \end{aligned} \tag{D.2}$$

We now average $\Delta D_a(n)$ over the different choices of the particular active leave that has been branched. This amounts to replacing d_a in the above formula by $\langle d_a\rangle_n$, the average depth of the *active* leaves in a tree of n leaves. Writing $D_i(n+1) = D_i(n) + \Delta D_i(n)$, for $i = a, b$, one would get a closed system for the quantities $D_i(n)$ provided $\langle d_a\rangle_n$ is expressed in terms of them. This can be done by

writing $\langle d_a \rangle_n = D_a(n)/a(n)$, where $a(n)$ is the expected number of active leaves in a tree of n leaves. This expected value is used here as an approximation to the actual number of active leaves.

The recurrence equations for $D_i(n)$ are specially simple in the most interesting case $p = 1/2$, since the dependence in $\langle d_a \rangle_n$ disappears from one of the equations:

$$D_a(n+1) = D_a(n) + 1 \tag{D.3}$$
$$D_b(n+1) = D_b(n) + \langle d_a \rangle_n + 1 \,. \tag{D.4}$$

The solution (with initial condition $D_a(1) = 0$) of Eq. (D.3) is simply:

$$D_a(n) = n - 1 \,. \tag{D.5}$$

Since the probabilities of an increment or decrement (by one unit) of the number of active leaves are the same and time-independent for $p = 1/2$, the number of active nodes performs a symmetric random walk with a reflecting boundary at 0 (this last condition arises from the prescription of setting active one node when the number of active nodes has reached zero in the previous step). For such random walk the expected value of active leaves, $a(n)$, increases as the square root of the number of steps. Since a new leaf is added at each time step, this leads to:

$$a(n) \sim n^{1/2} \,. \tag{D.6}$$

Combining (D.5) and (D.6) we obtain the average distance of active nodes from root at large tree sizes:

$$\langle d_a \rangle_n \approx \frac{D_a(n)}{a(n)} \sim n^{1/2} \,. \tag{D.7}$$

Now we can plug this result into Eq. (D.4), which can be solved recursively:

$$D_b(n) = D_b(1) + \sum_{t=1}^{n-1} (\langle d_a \rangle_t + 1) \sim \sum_{t=1}^{n-1} t^{1/2} \sim n^{3/2} \,. \tag{D.8}$$

The totally averaged depth, $\overline{N}_n$, which counts both the active and the inactive leaves, is

$$\overline{N}_n = \frac{D_a(n) + D_b(n)}{n} \sim \frac{n^{1/2} + n^{3/2}}{n} \sim n^{1/2} \,, \tag{D.9}$$

which explains the asymptotic behavior observed in Figure D.1 for $p = 1/2$.

We note that the growth dynamics presented here may be mapped to a branching process (Harris, 1963), with the difference that here the death (inactivation) of a node does not lead to its removal from the tree. The special case $p = 1/2$ corresponds to a critical branching process.

D.1 Discussion

As we have commented on in Chapter 5, activity model constitutes, together with Ford's alpha model, two simple models which lead to non-logarithmic scaling of the tree depth. In contrast with many of the available models having this behavior (Banavar et al., 1999; Aldous, 2001; Blum and François, 2006; Ford, 2006) activity model is formulated as a *dynamical* model, involving *growing trees*, so that rules are given to obtain the tree at the next time step from the present state. Its introduction has been motivated since, as we have

described in Chapter 3, one of the depth scaling behaviors suggested to explain the branching mechanisms of real phylogenies is a non-logarithmic scaling of the depth, such as: $d \sim n^{0.44}$ (remind that the scaling of d and $\overline{N}$ with n is the same).

As we discussed in Chapter 5, a recent analysis of several evolutionary models including species competition (Stich and Manrubia, 2009) indicates that in these models correlations are finally destroyed by mutation processes and persist only for a finite correlation time. Thus, sufficiently large trees would have a scaling behavior closer to the asymptotic ERM predictions. As described in Chapter 5, this finite-size transient behavior is also obtained with the evolvability model. Since the largest phylogenies in databases such as TreeBASE and PANDIT have only hundreds or some thousands of leaves, respectively, it is possible that the observed imbalance and depth scaling is a finite-size regime. Nevertheless, models going beyond the ERM scaling are needed at least to explain this finite-size regime, and also to elucidate the true asymptotic scaling behavior.

The final aim of the modeling of phylogenetic trees is to provide biological mechanisms explaining the branching topology of the Tree of Life. In this direction, the branching of internal edges in the Ford model (described in Section 2.3.2) has no obvious biological interpretation. The activity model puts the mechanisms of birth-death critical branching (Harris, 1963) within a framework of transitions between node internal states similar in spirit to the approach of Pinelis (2003). The need to tune a parameter to attain the non-ERM critical behavior is, however, a limitation for its applicability. Much additional work is still needed to identify the proper biological mechanisms behind evolutionary branching and adequate modeling of them.

Appendix E

Evolvability model with refractory period and mass extinctions

In Chapter 5 we proposed an evolutionary branching model based on the relevance of the existence of taxa which are unable to undergo a new diversification event. With the aim of exploring the biological meaning of the prevalence of these taxa in real phylogenies, in this appendix we are going to consider how the depth scaling of our evolvability model is affected by, on the one hand, the presence of a refractory period, i.e. this period of time in which a newly formed species is unable to speciate again (Chan and Moore, 1999), and, on the other hand, the presence of random mass extinction events.

As described in Chapter 5, the evolvability model is based on two possible outcomes. On the one hand, the new species inherit the same capacity as the mother species to speciate again. On the other hand, one of the daughter species is unable to speciate again, i.e. it cannot undergo a new diversification event. With the purpose of an-

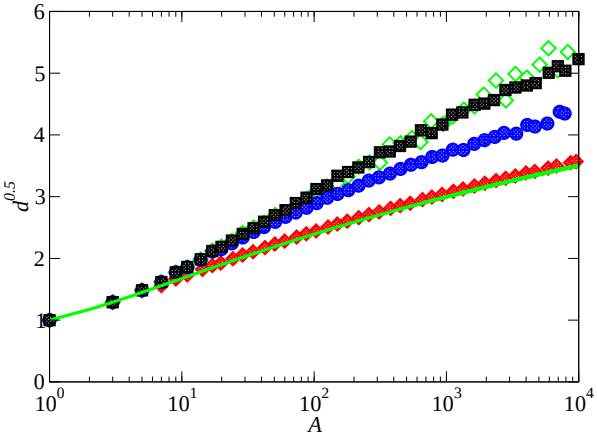


Figure E.1: Depth scaling of the evolvability model with refractory period. Mean depth scaling of the trees generated with the evolvability model for $p = 0.25$ (green empty diamonds) when adding different refractory periods: $RP = 1$ (red patterned diamonds) and $RP = 23$ (black patterned squares) for a total of 32 speciation events, and $RP = 8$ (blue patterned circles) for a total of 26 speciation events. Green line corresponds to the ERM model.

alyzing the effect of the refractory period on the depth scaling of the evolvability model, we replace the never-speciating outcome with a certain refractory period, that is, a certain period of time that the new species uncapable to speciate has to wait before being able to speciate again. In that sense, we included the refractory period (RP) in the generation of trees with $p = 0.25$, the ones which reproduced the depth scaling of the real phylogenies (Figure E.1). Those trees with $RP = 1$, i.e. one markovian step uncapable to speciate, reproduce the ERM-depth scaling and, as we were increasing the refractory period, the depth scaling was drawing away from the ERM-depth scaling, getting closer to the one described by the trees generated with the evolvability model without refractory period ($RP = \infty$) with $p = 0.25$. Therefore, for long refractory period, the depth scaling of the generated trees reproduced the same depth scaling of those trees with taxa unable to diversify with $p = 0.25$.

From the macroevolutionary point of view, if we consider a taxon that has suffered a mass extinction event as this in which all its daughter subtaxa were exterminated, from a topological point of view we could identify this taxon with a never-speciating taxon. In that sense, in order to analyze the consequences of mass extinction events in the evolvability model, as well as to prove if the mass extinction events can be considered as one of the biological causes of the existence of never-speciating taxa, we applied random mass extinction events to trees generated with the evolvability model where asymmetric and symmetric events had the same probability to take place ($p = 0.5$). For this purpose, we made a random selection of internal nodes of the phylogenetic trees, and eliminated all the subtrees pending from the selected one. After eliminating around 80% of the nodes of the phylogenetic tree with a $p = 0.5$, we obtained a depth scaling behavior similar to the one that fits with the phylogenetic trees without extinctions, that is with $p = 0.25$ (see Figure E.2). In a previous paper, Heard and Mooers (2002) reported that the presence of mass extinction events does not change the scaling of the ERM trees. We see here that, in the range of sizes in which our

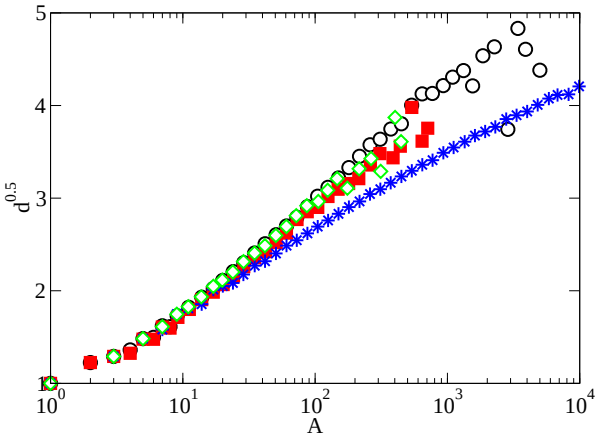


Figure E.2: Effect of the extinction events on the depth scaling of the evolvability model. The depth scaling of the trees generated for $p = 0.5$ (blue stars) and subjected to massive extinction events (eliminating the 80% of the nodes) (green empty diamonds) fits the depth scaling of proteins (red solid squares) and organisms (black empty circles).

model has non-ERM behavior, mass extinction events are able to alter the topology of the resulting trees.

In this appendix we have analyzed the effect of the existence of refractory period and mass extinction events over the depth scaling of the trees generated with evolvability model. This study proposes that, from the biological point of view, long refractory periods and mass extinction events can be considered as some of the evolutionary processes that can lead to the existence of those sorts of taxa that we identified here as taxa incapable of undergoing a new diversification event for a very long period of time.

Appendix F

Organism vs language taxonomies

In Chapter 6, with the main aim of characterizing the perturbing effect of the rank-based taxonomic criteria on the topological properties of the evolutionary trees, we described a comparative analysis between biological taxonomic and phylogenetic trees. In order to expand on the study of this phenomenon, in the present appendix we apply the same comparative approach to a different evolutionary system: language evolutionary trees. Thus, we will show the results from the comparative analysis between the trees obtained from a rank-based hierarchical classification of language families, and the phylogenetic trees of language families .

Firstly, for the analysis of language rank-based taxonomy, we resort to the database Ethnologue (2010), a catalogue of the 6,909 known living languages of the whole world, grouped in 128 families. Owing to the fact that the downloading of the whole database implied several problems derived from the way in which some classification ranks are labelled, we focused our analysis on the first five levels (starting from the root) out of the nine levels that the classification

Austronesian	Gray and Jordan (2000)
Bantu	Rexová et al. (2006)
Indo-European	Gray and Atkinson (2003)

Table F.1: Language phylogeny datasets.

can reach in the largely represented families, such as Niger-Congo (1532 languages) and Austronesian (1257 languages) language families. Thus, on March 15th 2010 we downloaded the set of 128 trees of the language families that constitute the Ethnologue database. Secondly, concerning the analysis of language phylogenies, the low number of works published on the application of the phylogenetic approach to the reconstruction of language evolution made it difficult to compile a representative database of language phylogenies. In this way, for our analysis, we compiled the 3 major works on the phylogenetic reconstruction of the language families which are most thoroughly studied (Austronesian, Bantu and Indo-European language families) (see Table F.1).

As was the case for organisms, the scaling of the mean depth, d, as function of the tree size of language taxonomy and language phylogenies differs. In this sense, it is remarkable that, while language phylogenies follow the same depth scaling as the one described for organism phylogenies (see Figure F.1), language rank-based taxonomy Ethnologue shows, like for the organism rank-based taxonomy CoL, a depth scaling behavior very similar to the one described by the fully polytomic tree, with a maximum mean depth of 5 for the largest language families.

Following the same methodology as when studying organism evolutionary trees, we used the degree distribution as a way to characterize the distribution of polytomies in the language taxonomy from Ethnologue. This distribution follows, as was the case of the rank-free taxonomy of organisms (ToL), a power-law scaling with

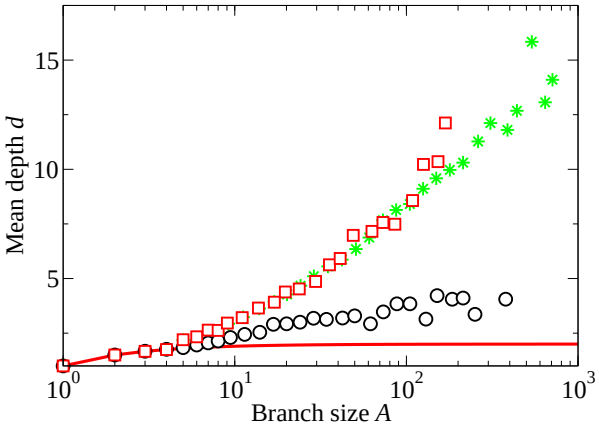


Figure F.1: Depth scaling of language taxonomies and language phylogenies. Plot of the logarithmically binned set of the mean depth scaling of Ethnologue (black empty circles) and language phylogenies (red empty squares) compared with the mean depth scaling of the organism phylogenies from TreeBASE (green stars). Red line represents the mean depth scaling of the fully polytomic tree.

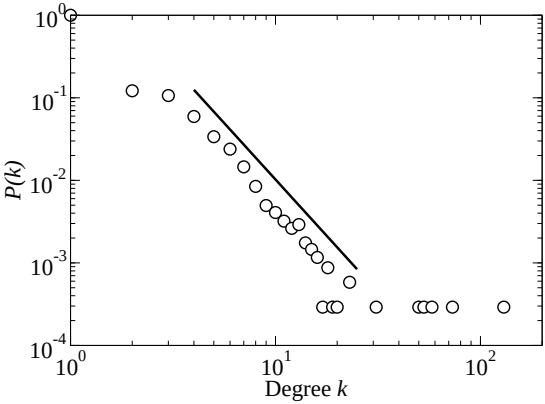


Figure F.2: Degree distribution of language taxonomic trees. Plot of the degree distribution of the language taxonomic trees obtained from Ethnologue. Black line corresponds to a power law $P(k) \sim k^{\gamma}$ with an exponent $\gamma = 2.73 \pm 0.07$.

the form $P(k) \sim k^{\gamma}$ (Figure F.2), with $\gamma \sim 2.73$. Note that, since the language phylogenies analyzed are almost fully binary, the degree distribution analysis of those phylogenies become irrelevant.

The degree-depth correlation analysis for language taxonomy and language phylogenies shows the same features as that for organism analysis described in Chapter 6. On the one hand, language phylogenetic trees show no correlation between the degree of the nodes and the average depth, with an average depth value for all the internal nodes of 3, the same value found for TreeBASE phylogenies. On the other hand, language taxonomy shows a correlation pattern, with a substantial presence of polytomies at all levels (Figure F.3).

In this appendix, with the main goal of characterizing the effect of the rank-based taxonomies on the topology of the language evolutionary trees, we exported the approach carried out for organisms in

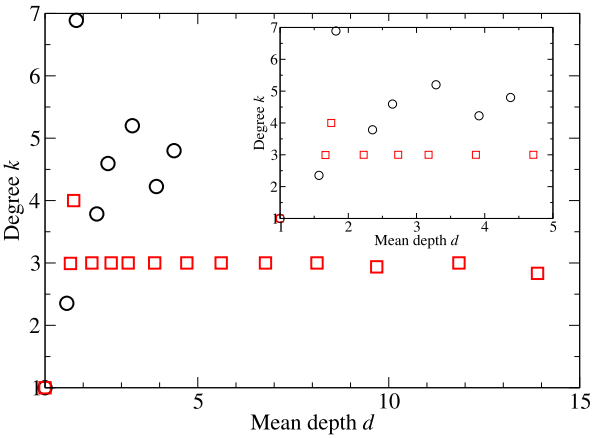


Figure F.3: Degree-depth correlation in language taxonomic and phylogenetic trees. Comparison between the degree-depth correlation of language taxonomic tree (black empty circles) and language phylogenies (red empty squares). Inset: Zoom of the degree-depth correlation of language taxonomic tree (black empty circles) and language phylogenies (red empty squares) for the range of average depth from 1-5.

Chapter 6 and applied it to the case of languages. We have described interesting similarities between language and organism phylogenies, and we also have observed that language rank-based taxonomy, represented by the Ethnologue database, showed an increase in the presence of polytomies with respect to phylogenies, which were almost fully binary. This increase is translated in a change in the depth scaling, and in the correlation exhibited between the degree and the average depth of the language taxonomy. The results reported here suggest that the disturbing effect of the rank-based taxonomic criteria over the branching pattern of the evolutionary trees, studied in Chapter 6, can be also found in language evolutionary trees. But wider analyses of a refined version of the Ethnologue dataset, as well as of other databases, such as the Multitree (2010) are needed in order to reach further conclusions.

Appendix G

Depth scaling measures for phylograms

The measures that we used in Chapters 3, 4, 5 and 6 for the depth scaling analysis were proposed in Section 2.2.3 for unweighted directed tree-like networks. A direct generalization of this approach is the definition of those measures for weighted directed tree-like networks. In this direction, Zhang (2009) and Zhang and Guo (2010) derived the branch size and the cumulative branch size for weighted networks. With the same purpose, in this appendix we will define branch size and cumulative branch size, for weighted tree-like networks, and apply these measures so as to characterize the depth scaling of the Timetree of Life used in Chapter 7.

In Section 2.2.3 we defined branch size, A_i, as follows: for a certain subtree rooted in node i, S_i, its branch size, A_i, is the number of subtaxa that diversify from node i (including itself). Likewise, its cumulative branch size, C_i , is the sum of the branch sizes associated to all the nodes in subtree S_i, $C_i = \sum A_j$. From these definitions we obtained respective measures for weighted tree-like networks. In a weighted subtree S_i, rooted in node i, the links are associated to

a scalar, l_{ij}, which in the case of phylograms represents the length of the branch that connects taxon i with its subtaxon j. The branch weight, A'_i, constitutes the sum of the lengths, l_{ij}, of all the links inside subtree S_i, and the weighted cumulative branch size, C'_i, would be defined as the sum of branch sizes A_j for all the nodes in subtree S_i, ponderated by its corresponding branch length, l_{ij}:

$$C'_i = \sum_j (A_j l_{ij}) \, .$$

As an expansion of these measures, we propose, firstly, a measure of the mean diversification time of a certain lineage by means of the *mean branch weight*

$$\langle A'_i \rangle = \frac{A'_i}{A_i - 1} \, ,$$

where $A_i - 1$ corresponds to the number of links in subtree S_i of this lineage. Secondly, we also propose a measure of the average depth in time units of a certain lineage through the *mean weighted depth*

$$\langle C'_i \rangle = \frac{C'_i}{A_i - 1} \, .$$

As observed in Figure G.1(a), the mean diversification time, $\langle A' \rangle$, as function of the size of subtree S_i, A_i, of Timetree does not display any correlation ($p = 0.38$, for non-significant correlation), showing a constant mean speciation time of around 100 million years (My). By comparing the scaling behavior of the mean weighted depth, $\langle C' \rangle$, with the scaling behavior of the measure for the mean depth, d, (used in Chapters 3, 4, 5, 6) we observe that the scaling of $\langle C' \rangle$ as function of the subtree size, A, of Timetree shows a scaling slope similar to the one displayed in the mean depth scaling, d, as function of the subtree size, A, (Figure G.1(b)).

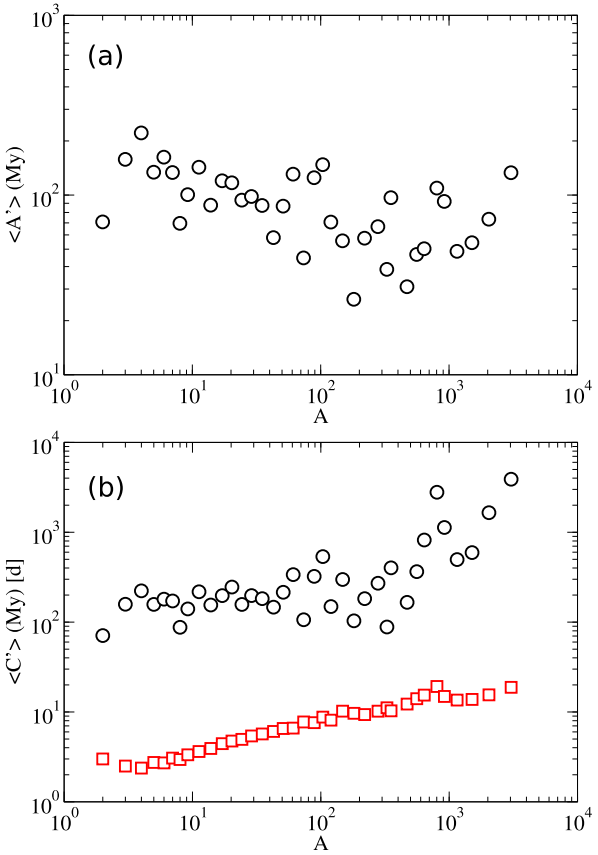


Figure G.1: Weighted depth scaling. (a) Logarithmically binned set of values of the mean branch weight (in million years), $\langle A' \rangle$, as function of the branch size, A, of Timetree. (b) Logarithmically binned set of values of the mean weighted depth (in million years), $\langle C' \rangle$, (black empty circles) and the mean depth, d, (red empty squares) as functions of the branch size, A, of Timetree.

APPENDIX G. DEPTH SCALING MEASURES FOR PHYLOGRAMS

The absence of correlation of the mean branch weight, $\langle A' \rangle$, with the size of the lineage, as well as the absence of divergence between the mean weighted depth and the mean depth scalings imply that neither $\langle A' \rangle$ nor $\langle C' \rangle$ measures provide additional information on branch length distribution which was not already in the topological analysis.

In this appendix we have proposed an extended definition of the branch size, A, and the cumulative branch size, C, for the case of weighted tree-like networks, mean branch weight, $\langle A' \rangle$, and mean weighted depth, $\langle C' \rangle$, respectively. Our findings show that these measures display no additional sensitivity in the characterization of the branch length distribution all over the Timetree. Therefore, we must continue the search for alternative definitions of these measures with the aim of applying the depth scaling approach for the characterization of the branch length distribution in phylograms.

Appendix H

Python codes

In this appendix we include the Python codes used for the computation of the depth scaling analysis (Section H.1), for the conversion of tree files from Newick format to columns format (Section H.2), as well as the Python code used for the simulation of the evolvability model (Section H.3).

H.1

AC.py

```
"""It computes A, C, A' and C' for each node of the phylo-
genetic tree. The input file is a .txt file where the phy-
logenetic tree is defined in columns format, and the output
file is a .dat file with 5 columns corresponding to: node,
A, C, A' and C'.

Copyleft 2010, Adrian Jacobo and Alejandro Herrada."""

import sys
```

```
import os

nombreArch = "prueba_tree.txt"

if len(sys.argv)>1:
  nombreArch = sys.argv[1]

lista = [ i for i in os.listdir("./") if ".txt" in i]

for nombreArch in lista:
  node1=[]
  node2=[]
  dist=[]
  nodes_all=[]
  hijosdelpadre={}
  hijos=[]
  j=0
  dic_dist={}
  a={}
  c={}
  a_d={}
  c_d={}
  flag=1
  auxleft=[]
  calcc=0
  tempc=0
  tempa=0
  calcc_d=0
  tempc_d=0
  tempa_d=0

  try:
    f = open(nombreArch ,"r")
    o = open("./out/"+nombreArch[0:-3]+'dat',"w")
```

```
except:
  print "El fichero '%s' no se pudo abrir"%nombreArch
  sys.exit()

for i in f:
  nodes = i.split()
  node1.append(int(nodes[0]))
  node2.append(int(nodes[1]))
  dist.append(float(nodes[2]))
nodes_all = [node1,node2,dist]

for i in nodes_all[1]:
  j=0
  while j < len(nodes_all[1]):
    if nodes_all[1][j] == i:
      hijos.append(nodes_all[0][j])
    j+=1
  hijosdelpadre[i]= hijos

j=0
for i in nodes_all[0]:
  dic_dist[i]=dist[j]
  j=j+1

for i in nodes_all[0]:
  if not hijosdelpadre.has_key(i):
    a[i]=1
    c[i]=1
    a_d[i]=0
    c_d[i]=0

flag=1
auxleft=hijosdelpadre.keys()
while flag==1:
  flag=0
```

```
    left=auxleft
    auxleft=[]
    for i in left:
      calcc=0
      tempc=0
      tempa=0
      calcc_d=0
      tempc_d=0
      tempa_d=0
      for j in hijosdelpadre[i]:
        if c.has_key(j):
          tempc=tempc+c[j]
          tempa=tempa+a[j]
        if c_d.has_key(j):
          tempc_d=tempc_d+a[j]*dic_dist[j]+c_d[j]
          tempa_d=tempa_d+a_d[j]+dic_dist[j]
        else:
          calcc=1
          calcc_d=1
          flag=1
          auxleft.append(i)
          break

      if calcc==0:
        c[i]=tempc+tempa+1
        a[i]=tempa+1
      if calcc_d==0:
        c_d[i]=tempc_d
        a_d[i]=tempa_d

  for i in a_d.keys():
    o.write ("%d\t %d\t %d\t %.4f\t %.4f\n" %(i, a[i], c[i],
    a_d[i], c_d[i]))
  o.close()
```

H.2

newick2columns.py

```
"""It converts phylogenetic trees represented in Newick
format into columns format. The input file is a .nwk file
where the phylogenetic tree is defined in Newick format,
and the output file is a .txt file with the phylogenetic
tree defined in columns format.

Copyleft 2010,Adrian Jacobo and Alejandro Herrada"""

import sys
import os
import re

nombreArch = "default.dat"
if len(sys.argv)>1:
  nombreArch = sys.argv[1]

lista = [ i for i in os.listdir("./") if ".nwk" in i]

for nombreArch in lista:
  bracket_index=[]
  tree=()
  p=0
  pold=0
  nodo_num=0
  acumulador=1
  pos=[]
  i=0
  listree=[]
  j=0
  dic={}
  rev_index=[]
```

```
try:
  file = open(nombreArch ,"r")
  o = open("./out/"+nombreArch[0:-3]+'txt',"w")
except:
  print "El fichero '%s' no se pudo abrir"%nombreArch
  sys.exit()

tree=file.readline()
if tree!='':

  nodo_num=0
  acumulador=1
  pos=[]
  bracket_index=[]
  listree=[]
  i=0
  tokens=[')',',',':','(',';']

  for j in range(len(tree)):
    if  tree[j] in tokens:
      pos.append(j)

  j=0
  dic={}
  nodenum=0
  for i in range(len(pos)):
    t=tree[pos[i]]
    if t == ')' or t=='(' or t == ',':
      if tree[pos[i+1]] != '(':
        listree.append(t)
        listree.append(nodenum)
        nodenum=nodenum+1
      else:
        listree.append(t)
```

```
  elif t == ':':
    dic[nodenum-1]=tree[pos[i]+1:pos[i+1]]

newtree=''
for i in listree:
  newtree=newtree+str(i)

level=0
niveles={}
closepar=0
for token in listree:
  if token =='(':
    level=level+1

  if type(token).__name__=='int':
    try:
      niveles[level].append(token)
    except:
      niveles[level]=[token]
    if closepar==1:
      closepar=0
      for j in niveles[level+1]:
        try:
          o.write("%s\t %s\t %s\n" %(j, token, dic[j]))
        except:
          pass
      niveles[level+1]=[]

  if token ==')':
    level=level-1
    closepar=1

o.close()
```

H.3

evolvabilitymodel.py

```
"""It runs the evolvability model. It prints two different
output files per tree generated: i) a .dat file with the
tree generated, defined in columns format; and ii) a .dat
file with the sequences of the leaves of the tree generated.

Copyleft 2010, Alejandro Herrada."""

import sys
import os
import random

####FUNCION PRIMERA (MUTACION ALOPATRICA)#####
def mutacion_alo(DIC, PAD, NEW_SEQS):
  seq=DIC[PAD]
  tamseq=len(seq)
  contdor=max(DIC.keys())
  pos=random.randrange(0,tamseq) #Selec. posicion
  mut=random.randrange(0,4) #Generar mutacion

  seq_new=seq[:pos]+[mut]+seq[pos+1:]
  contdor+=1
  DIC[contdor]=seq_new #incluye en {num(h):sequ_i}
  hijo1=contdor #asigna numero al hijo
  NEW_SEQS.append(contdor)
  o.write ('%d\t %d\n' %(hijo1, PAD))
  contdor+=1
  DIC[contdor]=seq #madre cambia numero y pasa a hija
  hijo2=contdor
  NEW_SEQS.append(contdor)
  o.write ('%d\t %d\n' %(hijo2, PAD))
```

```
####FUNCION PRIMERA (MUTACION SIMPATRICA)#####
def mutacion_sim(DIC, PAD, NEW_SEQS):
  seq=DIC[PAD]
  tamseq=len(seq)
  contdor=max(DIC.keys())
  pos=random.randrange(0,tamseq) #Selec. posicion
  mut=random.randrange(0,4) #Generar mutacion

  seq_new=seq[:pos]+[mut]+seq[pos+1:]
  contdor+=1
  DIC[contdor]=seq_new #incluye en {num(h):sequ_i}
  hijo1=contdor #asigna numero al hijo
  NEW_SEQS.append(contdor)
  o.write ('%d\t %d\n' %(hijo1, PAD))
  contdor+=1
  DIC[contdor]=seq #madre cambia numero y pasa a hija
  hijo2=contdor
  o.write ('%d\t %d\n' %(hijo2, PAD))

outfiles=0
while outfiles < 20:
  ######Hacer una lista para las secuencias######
  seq=[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
  0,0,0,0,0,0,0,0]

  ##########HACER LOS DICCIONARIOS##############
  contdor=0
  new_seqs=[]
  NUM_SEQS={}
  NUM_SEQS[contdor]=seq
```

```
hojas=[]
hojas.append(contdor)
o=open('./out4__05/tree'+str(outfiles)+'.dat',"w")
s=open('./out4__05/sequ'+str(outfiles)+'.dat','w')

mutevent=0
while mutevent < 17: ####Mutaciones#####
  new_seqs=[] #lista actual de hojas
  for hoja in hojas: #cada sec. de lista
    padre=hoja
    spp=random.random()
    if spp>0.5:###ESPECIACION ALOPATRICA###
      mutacion_alo(NUM_SEQS, padre, new_seqs)
    else:###ESPECIACION SIMPATRICA###
      mutacion_sim(NUM_SEQS, padre, new_seqs)
  hojas=new_seqs
  mutevent+=1
o.close()

for hoja in hojas:
  if NUM_SEQS.has_key(hoja):
    sec=''
    for num in NUM_SEQS[hoja]:
      if num==0:
        sec=sec+'a'
      if num==1:
        sec=sec+'c'
      if num==2:
        sec=sec+'g'
      if num==3:
        sec=sec+'t'
    s.write ('>sequence_%s\n %s\n' %(hoja, sec))
s.close()
outfiles+=1
```

Appendix I

Related publications

Publications related with this thesis:

- Herrada, E.A., Tessone, C.J., Klemm, K., Eguíluz, V.M. Hernández-García, E. and Duarte, C.M. 2008. *Universal scaling in the branching of the Tree of Life*. PloS ONE, 3: e2757.
- Hernández-García, E., Tuğrul, M., Herrada, E.A., Eguíluz, V.M. and Klemm, K. 2010. *Simple models for scaling in phylogenetic trees*. Int. J. Bifurcat. Chaos, 20: 805-811.

Publications in proceedings:

- Hernández-García, E. Herrada, E. A., Rozenfeld, A. F., Tessone, C. J., Eguíluz, V. M., Duarte, C. M., Arnaud-Haond, S., Serrão, E. 2007. *Evolutionary and Ecogical Trees and Networks*. Nonequilibrium Statistical Mechanics And Nonlinear Physics: XV Conference on Nonequilibrium Statistical Mechanics and Nonlinear Physics, Ed. by O. Descalzi, O.A. Rosso and H.A. Larrondo. AIP Conference Proceedings Volume 913, American Institute of Physics (New York, 2007), pp. 78-83.

APPENDIX I. RELATED PUBLICATIONS

Publications in preparation:

- Herrada, E.A. et al. *Scaling properties of protein family phylogenies*.
- Herrada, E.A. et al. *Phylogenies vs taxonomies: A topological study*.
- Herrada, E.A. et al. *Branch length scaling in the evolutionary tree*.

Conference presentations:

- Oral communication *The shape of phylogenetic trees: From taxonomic trees to the Tree of Life*. Darwin09: 150 Years after Darwin: From Molecular Evolution to Language. Palma (Spain), 23-27 November 2009.
- Poster Communication *Scaling properties in protein evolution*. FISES2008, XV Reunión de Física Estadística. Salamanca, 27-29 March 2008.
- Oral communication *Scaling properties in the Tree of Life*. Workshop on Dynamics and Evolution of Biological and Social Networks, Palma de Mallorca (Spain), 18-20 February 2008.
- Oral communication *Topological diversity in phylogenies: microevolution vs macroevolution*. XVI Seminario de Genética de Poblaciones y Evolución, Sant Feliu de Guíxols (Spain), 15-18 November 2006.
- Oral communication *Scaling properties in the Tree of Life*. Workshop on Social and Ecological Networks, European Conference on Complex Systems (ECCS06), Oxford (United Kingdom), 28-29 September 2006.
- Poster Communication *Scaling properties of intraspecific and interspecific phylogenies in the Tree of Life*. 10th Evolutionary Biology Meeting, Marseilles (France), 20-22 September 2006.

- Poster Communication *Scaling properties of intraspecific and interspecific phylogenies in the Tree of Life*. FISES2006, XIV Reunión de Física Estadística, Granada (Spain), 14-16 September 2006.

List of Figures

List of Tables

LIST OF TABLES

Bibliography

Agapow, P.-M. and A. Purvis. 2002. Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. Syst Biol 51:866–872.

Albert, R. and A.-L. Barabási. 2002. Statistical mechanics of complex networks. Rev Mod Phys 74:47–97.

Aldous, D. J. 1995. Probability distributions on cladograms. Pages 1–18. Random Discrete Structures. Springer, Berlin.

Aldous, D. J. 2001. Stochastic models and descriptive statistics for phylogenetic trees from Yule to today. Stat Sci 16:23–34.

Almaas, E. 2007. Biological impacts and context of network theory. J Exp Biol 210:1548–1558.

Almaas, E., P. L. Krapivsky, and S. Redner. 2005. Statistics of weighted treelike networks. Phys Rev E 71:036124.

Altaba, C. R. 2009. Universal Artifacts Affect the Branching of Phylogenetic Trees, Not Universal Scaling Laws. PLoS ONE 4:e4611.

AmphibiaTree. 2010. http://amphibiatree.org/.

BIBLIOGRAPHY

Anderson, P. W. 1972. More is different. Science 177:393–396.

Anderson, S. 1974. Patterns of faunal evolution. Q Rev Biol 49:311–332.

Anderson, S. 1975. On the number of categories in biological classification. Amer. Mus. Novitates 2584:1–9.

Aristóteles. 1994. Metafísica. Gredos, Madrid.

Assembling_the_Fungal_Tree_of_Life. 2010. http://aftol.org/.

Avery, O. T., C. M. Macleod, and M. McCarty. 1944. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. J Exp Med 79:137–158.

Avise, J. C. 1994. Molecular Markers, Natural History, and Evolution. Chapman & Hall, New York.

Avise, J. C. 2009. Timetrees: beyond cladograms, phenograms, and phylograms. Pages 19–25. The timetree of life. Oxford University Press, Oxford.

Avise, J. C. and D. Mitchell. 2007. Time to standardize taxonomies. Syst Biol 56:130–133.

Babushok, D. V., E. M. Ostertag, and H. H. Kazazian. 2007. Current topics in genome evolution: molecular mechanisms of new gene formation. Cell Mol Life Sci 64:542–554.

Balch, W. E., L. J. Magrum, G. E. Fox, R. S. Wolfe, and C. R. Woese. 1977. An ancient divergence among the bacteria. J Mol Evol 9:305–311.

Baldwin, B. G. and M. J. Sanderson. 1998. Age and rate of diversification of the Hawaiian silversword alliance (Compositae). Proc Natl Acad Sci U S A 95:9402–9406.

Baldwin, E. 1937. An Introduction to Comparative Biochemistry. Cambridge University Press, Cambridge.

Ball, P. 2009. Branches. Nature's patterns: A Tapestry in Three Parts. Oxford University Press, New York.

Banavar, J. R., J. Damuth, A. Maritan, and A. Rinaldo. 2002. Supply-demand balance and metabolic scaling. Proc Natl Acad Sci U S A 99:10506–10509.

Banavar, J. R., J. Damuth, A. Maritan, and A. Rinaldo. 2006. Comment on "Revising the distributive networks models of West, Brown and Enquist (1997) and Banavar, Maritan and Rinaldo (1999): Metabolic inequity of living tissues provides clues for the observed allometric scaling rules" by Makarieva, Gorshkov and Li. J Theor Biol 239:391–393.

Banavar, J. R., A. Maritan, and A. Rinaldo. 1999. Size and form in efficient transportation networks. Nature 399:130–132.

Barabási, A.-L. and R. Albert. 1999. Emergence of scaling in random networks. Science 286:509–512.

Barabási, A.-L., R. Albert, and H. Jeong. 1999. Mean-field theory fro scale-free random networks. Physica A 272:173–187.

Barraclough, T. G. and S. Nee. 2001. Phylogenetics and speciation. Trends Ecol Evol 16:391–399.

Barthélemy, M. and A. Flammini. 2006. Optimal traffic networks. J Stat Mech 07:L07002.

Basilio, A. M., D. Medan, J. P. Torretta, and N. J. Bartoloni. 2006. A year-long plant-pollinator network. Austral Ecol 31:975–983.

Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. 2004. The Pfam protein families database. Nucleic Acids Res 32:D138–D141.

Benton, M. J. 2000. Stems, nodes, crown clades, and rank-free lists: is Linnaeus dead? Biol Rev Camb Philos Soc 75:633–648.

Berg, J., M. Lässig, and A. Wagner. 2004. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. BMC Evol Biol 4:51.

Bienaymé, I. J. 1845. De la loi de multiplication et de la durée des familles. Soc Philomath Paris Extraits 5:37–39.

Bisby, F. A., Y. R. Roskov, T. M. Orrell, D. Nicolson, L. E. Paglinawan, N. Bailly, P. M. Kirk, T. Bourgoin, and G. Baillargeon, eds. 2010. Species 2000 & ITIS Catalogue of Life: 2010 Annual Checklist Taxonomic Classification. DVD. Species 2000, Reading, UK.

Blum, M. G. B. and O. François. 2005. On statistical tests of phylogenetic tree imbalance: the Sackin and other indices revisited. Math Biosci 195:141–153.

Blum, M. G. B. and O. François. 2006. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. Syst Biol 55:685–691.

Blüthgen, N., F. Menzel, T. Hovestadt, and B. Fiala. 2007. Specialization, constraints, and conflicting interests in mutualistic networks. Curr Biol 17:341–346.

Blüthgen, N., D. Mezger, and K. E. Linsenmair. 2006. Ant-hemipteran trophobioses in Bornean rainforest – diversity, specificity and monopolisation. Insectes Soc 53:194–203.

Blüthgen, N., N. E. Stork, and K. Fiedler. 2004. Bottom-up control and co-occurrence in complex communities: honeydew and nectar determine rainforest ant mosaic. Oikos 106:344–358.

Boccaletti, S., V. Latora, and Y. Moreno, eds. 2010. Handbook of Biological Networks. vol. 10 of *World Scientific Lecture Notes in Complex Systems.* World Scientific Co. Pte. Ltd., Singapore.

Boguñá, M. and R. Pastor-Satorras. 2002. Epidemic spreading in correlated complex networks. Phys Rev E 66:047104.

Bollobás, B. and O. Riordan. 2004. Shortest paths and load scaling in scale-free trees. Phys Rev E 69:036114.

Bornholdt, S. and H. Ebel. 2001. World Wide Web scaling exponent from Simon's 1955 model. Phys Rev E 64:035104.

Boto, L. 2010. Horizontal gene transfer in evolution: facts and challenges. Proc Biol Sci 277:819–827.

Britten, R. J. 2006. Almost all human genes resulted from ancient duplication. Proc Natl Acad Sci U S A 103:19027–19032.

Bromham, L. and D. Penny. 2003. The modern molecular clock. Nat Rev Genet 4:216–224.

Brookfield, J. F. Y. 2009. Evolution and evolvability: celebrating Darwin 200. Biol Lett 5:44–46.

Brose, U., E. L. Berlow, and N. D. Martinez. 2005. Scaling up keystone effects from simple to complex ecological networks. Ecol Lett 8:1317–1325.

Brose, U., T. Jonsson, E. L. Berlow, P. Warren, C. Banasek-Richter, L.-F. Bersier, J. L. Blanchard, T. Brey, S. R. Carpenter, M.-F. C. Blandenier, L. Cushing, H. A. Dawah, T. Dell, F. Edwards, S. Harper-Smith, U. Jacob, M. E. Ledger, N. D. Martinez, J. Memmott, K. Mintenbeck, J. K. Pinnegar, B. C. Rall, T. S. Rayner, D. C. Reuman, L. Ruess, W. Ulrich, R. J. Williams, G. Woodward, and J. E. Cohen. 2006. Consumer-resource body-size relationships in natural food webs. Ecology 87:2411–2417.

Brown, J. H. 1995. Macroecology. University of Chicago Press, Chicago.

Brown, J. H., J. F. Gillooly, A. P. Allen, V. M. Savage, and G. B. West. 2004. Toward a metabolic theory of ecology. Ecology 85:1771–1789.

Brown, J. K. M. 1994. Probabilities of evolutionary trees. Syst Biol 43:78–91.

Bryant, H. N. and P. D. Cantino. 2002. A review of criticisms of phylogenetic nomenclature: is taxonomic freedom the fundamental issue? Biol Rev Camb Philos Soc 77:39–55.

Bukovinszky, T., F. J. F. van Veen, Y. Jongema, and M. Dicke. 2008. Direct and indirect effects of resource quality on food web structure. Science 319:804–807.

Bullmore, E. and O. Sporns. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci 10:186–198.

Burlando, B. 1990. The fractal dimension of taxonomic systems. J Theor Biol 146:99–114.

Burlando, B. 1993. The fractal geometry of evolution. J Theor Biol 163:161–172.

Burnett, J. 1974. Of the Origin and Progress of Language. Facsimile of 1773-92 ed. Georg Olms Verlag, Edinburgh.

Butlin, R., J. Bridle, and D. Schluter, eds. 2009. Speciation and Patterns of Diversity. Ecological Reviews. Cambridge University Press, Cambridge.

Cadotte, M. W., T. J. Davies, J. Regetz, S. W. Kembel, E. Cleland, and T. H. Oakley. 2010. Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. Ecol Lett 13:96–105.

Caldarelli, G., C. C. Cartozo, P. D. los Rios, and V. D. P. Servedio. 2004. Widespread occurrence of the inverse square distribution in social sciences and taxonomy. Phys Rev E 69:035101.

Camacho, J. and A. Arenas. 2005. Food-web topology: universal scaling in food-web structure? Nature 435:E3–4; discussion E4.

Campos, P. R. A. and V. M. de Oliveira. 2004. Emergence of allometric scaling in genealogical trees. Adv Complex Syst 7:39–46.

Cantino, P. D. and K. de Queiroz. 2000. PhyloCode: A Phylogenetic Code of Biological Nomenclature. PhyloCode.

Capocci, A., F. Rao, and G. Caldarelli. 2008. Taxonomy and clustering in collaborative systems: The case of the on-line encyclopedia Wikipedia. Europhys Lett 81:28006.

Carroll, S. B. 2005. Evolution at two levels: on genes and form. PLoS Biol 3:e245.

Cartozo, C. C., D. Garlaschelli, C. Ricotta, M. Barthélemy, and G. Caldarelli. 2008. Quantifying the taxonomic diversity in real species communities. J Phys A 41:224012.

Castelló, X. 2010. Collective phenomena in social dynamics: consensus problems, ordering dynamics and language competition. Ph.D. thesis Universitat de las Illes Balears.

Catalogue_of_Life. 2010. http://www.catalogueoflife.org/.

Cavalier-Smith, T. 2004. Only six kingdoms of life. Proc Biol Sci 271:1251–1262.

Cavalli-Sforza, L. L. and A. W. F. Edwards. 1967. Phylogenetic analysis: models and estimation procedures. Am J Hum Genet 19:233–257.

Cavender-Bares, J., K. H. Kozak, P. V. A. Fine, and S. W. Kembel. 2009. The merging of community ecology and phylogenetic biology. Ecol Lett 12:693–715.

Celko, J. 2004. Joe Celko's Trees and Hierarchies in SQL for Smarties. Morgan Kaufmann Publishers, San Francisco.

Chan, K. M. A. and B. R. Moore. 1999. Accounting for mode of speciation increases power and realism of tests of phylogenetic asymmetry. Am Nat 153:332–346.

Chan, K. M. A. and B. R. Moore. 2002. Whole-tree methods for detecting differential diversification rates. Syst Biol 51:855–865.

Chatton, É. 1925. Pansporella perplexa. Réflexions sur la biologie et la phylogénie des protozoaires. Ann Sci Nat Zool 10-VII:1–84.

Chothia, C. and J. Gough. 2009. Genomic and structural aspects of protein evolution. Biochem J 419:15–28.

Chu, J. and C. Adami. 1999. A simple explanation for taxon abundance patterns. Proc Natl Acad Sci U S A 96:15017–15019.

Colless, D. H. 1982. Phylogenetics: the theory and practice of phylogenetic systematics. Syst Zool 31:100–104.

Confederacion_Hidrografica_del_Mino-Sil. 2010. http://www.chminosil.es/.

Copeland, H. F. 1938. The kingdoms of organisms. Q Rev Biol 13:383–420.

Corbet, A. S. 1942. The distribution of butterflies in the Malay Peninsula. Proc R Entomol Soc Lond A 16:101–116.

Cotton, J. A. and R. D. M. Page. 2006. The shape of human gene family phylogenies. BMC Evol Biol 6:66.

Cracraft, J. and M. J. Donoghue. 2004. Assembling the Tree of Life. Oxford University Press, Oxford.

cubiFOR. 2010. http://www.cesefor.com/cubifor/.

Cuvier, G. and A. Brongniart, eds. 1822. Description géologique des environs de Paris. Dufour et E. d'Ocagne, Paris.

Cypriniformes_Tree_of_Life. 2010. http://bio.slu.edu/mayden/cypriniformes/home.html.

da F. Costa, L., F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. 2007. Characterization of complex networks: A survey of measurements. Adv Phys 56:167–242.

Daniels, B. C., Y.-J. Chen, J. P. Sethna, R. N. Gutenkunst, and C. R. Myers. 2008. Sloppiness, robustness, and evolvability in systems biology. Curr Opin Biotechnol 19:389–395.

Darwin, C. 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. John Murray, London.

Darwin, C. and A. R. Wallace. 1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. J Proc Linn Soc Lond Zool 3:46–50.

Darwin, E. 1794-1796. Zoonomia; or, the laws of organic life. J. Johnson, London.

Davies, T. J., V. Savolainen, M. W. Chase, P. Goldblatt, and T. G. Barraclough. 2005. Environment, area, and diversification in the species-rich flowering plant family Iridaceae. Am Nat 166:418–425.

Dawkins, R. 1989. The evolution of evolvability. Pages 201–220 *in* Artificial Life. The Proceedings of an Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems, Vol. VI, September 1987. (C. Langton, ed.) Addison-Wesley Pub. Corp., Los Alamos.

Dayhoff, M. O. 1965-1978. Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington.

De Los Rios, P. 2001. Power law size distribution of supercritical random trees. Europhys Lett 56:898–903.

de Queiroz, K. 1988. Systematics and the darwinian revolution. Phylos Sci 55:238–259.

de Queiroz, K. 1997. The linnaean hierarchy and the evolutionization of taxonomy, with emphasis on the problem of nomenclature. Aliso 15:125–144.

de Queiroz, K. 2005. Linnaean, rank-based, and phylogenetic nomenclature: Restoring primacy to the link between names and taxa. Symb Bot Ups 33:127–140.

Dial, K. P. and J. M. Marzluff. 1989. Nonrandom diversification within taxonomic assemblages. Syst Zool 38:26–37.

Dobzhansky, T. 1937. Genetics and the origin of species. Columbia University Press, New York.

Doolittle, W. F. 2000. Uprooting the tree of life. Sci Am 282:90–95.

Dubois, A. 2007. Naming taxa from cladograms: a cautionary tale. Mol Phylogenet Evol 42:317–330.

Dunne, J. A., R. J. Williams, and N. D. Martinez. 2002. Food-web structure and network theory: The role of connectance and size. Proc Natl Acad Sci U S A 99:12917–12922.

Durrett, R. 2007. Random Graph Dynamics. Cambridge University Press, Cambridge.

Dyer, R. J. 2007. The evolution of genetic topologies. Theor Popul Biol 71:71–79.

Dyer, R. J. and J. D. Nason. 2004. Population Graphs: the graph theoretic shape of genetic structure. Mol Ecol 13:1713–1727.

Early_Bird. 2010. http://www.fieldmuseum.org/research_collections-/zoology/zoo_sites/early_bird/.

Eck, R. V. and M. O. Dayhoff. 1966. Atlas of Protein Sequence and Structure. Silver Springs, Maryland.

Eguíluz, V. M., E. Hernández-García, O. Piro, and K. Klemm. 2003. Effective dimensions and percolation in hierarchically structured scale-free networks. Phys Rev E 68:055102.

Eldredge, N. and S. J. Gould. 1972. Punctuated equilibria: An alternative to phyletic gradualism. Pages 82–115. Models in Paleobiology. Freeman, Cooper and Company, San Francisco.

Erdös, P. and A. Rényi. 1959. On Random Graphs. I. Pub Math 6:290–297.

Ereshefsky, M. 2007. Foundational issues concerning taxa and taxon names. Syst Biol 56:295–301.

Erwin, D. H. 2000. Macroevolution is more than repeated rounds of microevolution. Evol Dev 2:78–84.

Ethnologue. 2010. http://www.ethnologue.com/.

Euler, L. 1741. Solutio problematis ad geometriam situs pertinentis. Comment Acad Sci U Petrop 8:128–140.

Farris, J. S. 1973. On comparing the shape of taxonomic trees. Syst Zool 22:50–54.

Farris, J. S. 1976. Expected asymmetry of phylogenetic trees. Syst Zool 25:196–198.

Fauquet, C. M., M. A. Mayo, J. Maniloff, U. Desselberger, and L. A. Ball, eds. 2005. Virus Taxonomy: VIIIth Report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press, London.

Felsenstein, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst Zool 22:240–249.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376.

Fiala, K. L. and R. R. Sokal. 1985. Factors determining the accuracy of cladogram estimation: Evaluation using computer simulation. Evolution 39:609–622.

Filipchenko, Y. 1927. Variabilität und variation. Gebruder Bortraeger, Berlin.

Fisher, R. A. 1930. The genetical theory of natural selection. Clarendon Press, Oxford.

Fisher, R. A., A. S. Corbet, and C. B. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. J Anim Ecol 12:42–58.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. Syst Zool 19:99–106.

Fitch, W. M. 1971. Rate of change of concomitantly variable codons. J Mol Evol 1:84–96.

Fitch, W. M. 1977. On the problem of discovering the most parsimonious tree. Am Nat 111:223–257.

Fitch, W. M. and E. Margoliash. 1967. Construction of phylogenetic trees. Science 155:279–284.

Fitch, W. M. and E. Markowitz. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4:579–593.

Flessa, K. W. and R. H. Thomas. 1985. Modeling the biogeographic regulation of evolutionary rates. Pages 355–376. Phanerozoic diversity patterns: Profiles in macroevolution. Princeton University Press, Princeton.

FLYTREE. 2010. http://www.inhs.illinois.edu/research/FLYTREE/.

Fontdevila, A. and A. Moya. 2003. Evolución : origen, adaptación y divergencia de las especies. Síntesis, Madrid.

Ford, D. J. 2006. Probabilities on cladograms:introduction to the alpha model. Ph.D. thesis Stanford University Stanford.

François, O. and C. Mioland. 2007. Gaussian approximations for phylogenetic branch length statistics under stochastic models of biodiversity. Math Biosci 209:108–123.

Franklin, R. 1952. Photo 51.

Fraser, H. B. 2005. Modularity and evolutionary constraint on proteins. Nat Genet 37:351–352.

Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman. 2002. Evolutionary rate in the protein interaction network. Science 296:750–752.

Freeman, S. and J. C. Herron. 2001. Evolutionary analysis. Prentice-Hall Inc., Upper Saddle River, New Jersey.

Fuchs, J., E. Pasquet, A. Couloux, J. Fjeldså, and R. C. K. Bowie. 2009. A new Indo-Malayan member of the Stenostiridae (Aves: Passeriformes) revealed by multilocus sequence data: biogeographical implications for a morphologically diverse clade of flycatchers. Mol Phylogenet Evol 53:384–393.

Fuchs, J., J.-M. Pons, S. M. Goodman, V. Bretagnolle, M. Melo, R. C. K. Bowie, D. Currie, R. Safford, M. Z. Virani, S. Thomsett, A. Hija, C. Cruaud, and E. Pasquet. 2008. Tracing the colonization history of the Indian Ocean scops-owls (Strigiformes: Otus) with further insight into the spatio-temporal origin of the Malagasy avifauna. BMC Evol Biol 8:197.

Fusco, G. and Q. C. Cronk. 1995. A new method for evaluating the shape of large phylogenies. J Theor Biol 175:235–243.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18:866–873.

Galton, F. and H. W. Watson. 1874. On the probability of the extinction of families. J R Anthropol Inst 4:138–144.

Garlaschelli, D., G. Caldarelli, and L. Pietronero. 2003. Universal scaling relations in food webs. Nature 423:165–168.

Garrick, R. C., A. Caccone, and P. Sunnucks. 2010. Inference of population history by coupling exploratory and model-driven phylogeographic analyses. Int J Mol Sci 11:1190–1227.

Garrick, R. C., J. D. Nason, C. A. Meadows, and R. J. Dyer. 2009. Not just vicariance: phylogeography of a Sonoran Desert euphorb indicates a major role of range expansion along the Baja peninsula. Mol Ecol 18:1916–1931.

Gauthier, J. 1986. Saurischian Monophyly and the Origin of Birds. Mem Cal Acad Sci 8:1–55.

Gavrilets, S. 2003. Perspective: models of speciation: what have we learned in 40 years? Evolution 57:2197–2215.

Ghim, C.-M., K.-I. Goh, and B. Kahng. 2005. Lethality and synthetic lethality in the genome-wide metabolic network of Escherichia coli. J Theor Biol 237:401–411.

Ghim, C.-M., E. Oh, K.-I. Goh, B. Kahng, and D. Kim. 2004. Packet transport along the shortest pathways in scale-free networks. Eur Phys J B 38:193–199.

Giordano, A. R., B. J. Ridenhour, and A. Storfer. 2007. The influence of altitude and topography on genetic structure in the long-toed salamander (Ambystoma macrodactulym). Mol Ecol 16:1625–1637.

Gogvadze, E. and A. Buzdin. 2009. Retroelements and their impact on genome evolution and functioning. Cell Mol Life Sci 66:3727–3742.

Grantham, T. 2007. Is macroevolution more than succesive rounds of microevolution? Paleontology 50:75–85.

Grasa Hernández, R. 2002. El evolucionismo : de Darwin a la sociobiología. Ediciones Pedagógicas, Madrid.

GreenPhylDB. 2010. http://greenphyl.cirad.fr/cgi-bin/greenphyl.cgi.

Gregory, T. R. 2008. Understanding Evolutionary Trees. Evo Edu Outreach 1:121–137.

Griffiths, A. J. F., J. H. Miller, D. T. Suzuki, R. C. Lewontin, and W. M. Gelbart. 2000. An introduction to genetic analysis. W. H. Freeman, New York.

Guthrie, W. K. C. 1962. I. The earlier Presocratics and the Pythagoreans. A History of Greek Philosophy. Cambridge University Press, Cambridge.

Guyer, C. and J. B. Slowinski. 1991. Comparisons between observed phylogenetic topologies with null expectation among three monophyletic lineages. Evolution 45:340–350.

Guyer, C. and J. B. Slowinski. 1993. Adaptive radiation an the topology of large phylogenies. Evolution 47:253–263.

Haeckel, E. 1866. Generelle Morphologie der Organismen : allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von C. Darwin reformirte Decendenz-Theorie. Reimer, Berlin.

Hahn, M. W., G. C. Conant, and A. Wagner. 2004. Molecular evolution in large genetic networks: does connectivity equal constraint? J Mol Evol 58:203–211.

Hahn, M. W. and A. D. Kern. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22:803–806.

Haldane, J. B. S. 1932. The causes of evolution. Longman Green, London.

Halliburton, R. 2004. Introduction to population genetics. Pearson Education, Upper Saddle River, New Jersey.

Harding, E. F. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. Adv Appl Prob 3:44–77.

Hardy, G. H. 1908. Mendelian proportions in a mixed population. Science 28:49–50.

Harris, T. E. 1963. The theory of branching processes. Springer Verlag, Berlin.

Harrison, P. M. and M. Gerstein. 2002. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. J Mol Biol 318:1155–1174.

Hartmann, K., D. Wong, and T. Stadler. 2010. Sampling trees from evolutionary models. Syst Biol 59:465–476.

Harvey, P. H., R. K. Colwell, J. W. Silvertown, and R. M. May. 1983. Null models in ecology. Ann Rev Ecol Syst 14:189–211.

Harvey, P. H., R. M. May, and S. Nee. 1994. Phylogenies without fossils. Evolution 48:523–529.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174.

Heard, S. B. 1992. Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. Evolution 46:1818–1826.

Heard, S. B. and A. Ø. Mooers. 2002. Signatures of random and selective mass extinctions in phylogenetic tree balance. Syst Biol 51:889–897.

Hedges, S. B., J. Dudley, and S. Kumar. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22:2971–2972.

Hedges, S. B. and S. Kumar, eds. 2009. The timetree of life. Oxford University Press, Oxford.

Hennig, W. 1950. Grundzüge einer Theorie der phylogenetischen Systematik. Deutscher Zentralverlag, Berlin.

Hennig, W. 1966. Phylogenetic Systematics. University of Illinois Press, Urbana.

Hernández-García, E., M. Tuğrul, E. A. Herrada, V. M. Eguíluz, and K. Klemm. 2010. Simple models for scaling in phylogenetic trees. Int J Bifurcat Chaos 20:805–811.

Herrada, E. A., C. J. Tessone, K. Klemm, V. M. Eguíluz, E. Hernández-García, and C. M. Duarte. 2008. Universal Scaling in the Branching of the Tree of Life. PLoS ONE 3:e2757.

Hillis, D. M. 2007. Constraints in naming parts of the Tree of Life. Mol Phylogenet Evol 42:331–338.

Hillis, D. M., B. K. Mable, and C. Moritz. 1996. Applications of molecular systematics: The state of the field and a look to the future. Pages 515–543. Molecular Systematics. Sinauer Associates, Sunderland.

Hitchcock, E. 1840. Elementary Geology. J. S. & C. Adams, Amherst.

Ho, S. Y. W. and G. Larson. 2006. Molecular clocks: when times are a-changin'. Trends Genet 22:79–83.

HOVERGEN. 2010. http://pbil.univ-lyon1.fr/databases-/hovergen.php.

Hubbell, S. P. 2001. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press, Princeton.

Huelsenbeck, J. P. 2002. Testing a covariotide model of DNA substitution. Mol Biol Evol 19:698–707.

Huerta-Cepas, J., A. Bueno, J. Dopazo, and T. Gabaldón. 2008. PhylomeDB: a database for genome-wide collections of gene phylogenies. Nucleic Acids Res 36:D491–D496.

Hurst, L. D. 2009. Fundamental concepts in genetics: genetics and the understanding of selection. Nat Rev Genet 10:83–93.

Huynen, M. A. and E. van Nimwegen. 1998. The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol 15:583–589.

HymAToL. 2010. http://www.hymatol.org/.

Ings, T. C., J. M. Montoya, J. Bascompte, N. Blüthgen, L. Brown, C. F. Dormann, F. Edwards, D. Figueroa, U. Jacob, J. I. Jones, R. B. Lauridsen, M. E. Ledger, H. M. Lewis, J. M. Olesen, F. J. F. van Veen, P. H. Warren, and G. Woodward. 2009. Ecological networks: Beyond food webs. J Anim Ecol 78:253–269.

Jablonski, D. 1991. Extinctions: a paleontological perspective. Science 253:754–757.

Jeong, H., S. P. Mason, A. L. Barabási, and Z. N. Oltvai. 2001. Lethality and centrality in protein networks. Nature 411:41–42.

Jordan, I. K., Y. I. Wolf, and E. V. Koonin. 2003. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3:1.

Jordano, P., J. Bascompte, and J. M. Olesen. 2003. Invariant properties in coevolutionary networks of plant-animal interactions. Ecol Lett 6:69–81.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–123. Mammalian Protein Metabolism. Academic Press, New York.

Junker, B. H. and F. Schreiber, eds. 2008. Analysis of biological networks. john Wiley & Sons, Inc., New Jersey.

Kareiva, P. 2004. Ecology: Compensating for Extinction. Curr Biol 14:R627–R628.

Keller, R., R. Boyd, and Q. Wheeler. 2003. The illogical basis of phylogenetic nomenclature. Bot. Rev. 69:93–110.

Képès, F., ed. 2007. Biological networks. World Scientific Co. Pte. Ltd., Singapore.

Kim, P.-J., D.-Y. Lee, T. Y. Kim, K. H. Lee, H. Jeong, S. Y. Lee, and S. Park. 2007. Metabolite essentiality elucidates robustness of Escherichia coli metabolism. Proc Natl Acad Sci U S A 104:13638–13642.

Kimura, M. 1968. Evolutionary rate at the molecular level. Nature 217:624–626.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Kimura, M. and T. Ota. 1974. On some principles governing molecular evolution. Proc Natl Acad Sci U S A 71:2848–2852.

Kingman, J. F. C. 1982a. The coalescent. Stochastic Process Appl 13:235–248.

Kingman, J. F. C. 1982b. On the genealogy of large populations. J Appl Probab 19A:27–43.

Kirkpatrick, M. and M. Slatkin. 1993. Searching for Evolutionary Patterns in the Shape of a Phylogenetic Tree. Evolution 47:1171–1181.

Klemm, K., V. M. Eguíluz, and M. S. Miguel. 2005. Scaling in the structure of directory trees in a computer cluster. Phys Rev Lett 95:128701.

Kolaczkowski, B. and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980–984.

Koonin, E. V. 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39:309–338.

Koonin, E. V., Y. I. Wolf, and G. P. Karev. 2002. The structure of the protein universe and genome evolution. Nature 420:218–223.

Kreitman, M. 1996. The neutral theory is dead. Long live the neutral theory. Bioessays 18:678–683; discussion 683.

Kumar, S. 2005. Molecular clocks: four decades of evolution. Nat Rev Genet 6:654–662.

Kutschera, U. and K. J. Niklas. 2004. The modern theory of biological evolution: an expanded synthesis. Naturwissenschaften 91:255–276.

Kwoh, C. K. and P. Y. Ng. 2007. Network analysis approach for biology. Cell Mol Life Sci 64:1739–1751.

LaBarbera, M. 1989. Analyzing Body Size as a Factor in Ecology and Evolution. Annu Rev Ecol Syst 20:97–117.

Lamarck, J.-B. 1809. Philosophie zoologique. Paris.

Lapage, S. P., P. H. A. Sneath, E. F. Lessel, V. B. D. Skerman, H. P. R. Seeliger, and W. A. Clark, eds. 1992. International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision. American Society for Microbiology Press, Washington D. C.

Lartillot, N. and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109.

Lässig, M. and A. Valleriani, eds. 2002. Biological Evolution and Statistical Physics. Springer Verlag, Heidelberg.

Latora, V. and M. Marchiori. 2001. Efficient behavior of small-world networks. Phys Rev Lett 87:198701.

Lemey, P., M. Salemi, and A.-M. Vandamme, eds. 2009. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. Cambridge University Press, Cambridge.

Lenski, R. E., J. E. Barrick, and C. Ofria. 2006. Balancing robustness and evolvability. PLoS Biol 4:e428.

Leopold, L. B. 1971. Trees and Streams: The Efficiency of Branching Patterns. J Theor Biol 31:339–354.

Levchenko, A. 2001. Computational cell biology in the post-genomic era. Mol Biol Rep 28:83–89.

Lewis, O. T., J. Memmott, J. Lasalle, C. H. C. Lyal, C. Whitefoord, and H. C. J. Godfray. 2002. Structure of diverse tropical forest insect-parasitoid community. J Anim Ecol 71:855–873.

Li, H., A. Coghlan, J. Ruan, L. J. Coin, J.-K. Hériché, L. Osmotherly, R. Li, T. Liu, Z. Zhang, L. Bolund, G. K.-S. Wong, W. Zheng, P. Dehal, J. Wang, and R. Durbin. 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res 34:D572–D580.

Li, S. 1996. Phylogenetic Tree Construction Using Markov Chain Monte Carlo. Ph.D. thesis Ohio State University Columbus.

Li, W.-H. 1997. Molecular evolution. Sinauer Associates, Sunderland.

Li, W. L. S. and A. G. Rodrigo. 2009. Covariation of branch lengths in phylogenies of functionally related genes. PLoS One 4:e8487.

Lin, M., D. A. Payne, and J. R. Schwarz. 2003. Intraspecific diversity of Vibrio vulnificus in Galveston Bay water and oysters as determined by randomly amplified polymorphic DNA PCR. Appl Environ Microbiol 69:3170–3175.

Linder, H. P., P. Eldenäs, and B. G. Briggs. 2003. Contrasting patterns of radiation in African and Australian Restionaceae. Evolution 57:2688–2702.

Luscombe, N. M., J. Qian, Z. Zhang, T. Johnson, and M. Gerstein. 2002. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. Genome Biol 3:research0040.

Maddison, D. R., K.-S. Schulz, and W. P. Maddison. 2007. The Tree of Life Web Project. Zootaxa 1668:19–40.

Maddison, W. P. 1989. Reconstructing character evolution on polytomous cladograms. Cladistics 5:365–377.

Maddison, W. P. 2006. Confounding asymmetries in evolutionary diversification and character change. Evolution 60:1743–1746.

Magurran, A. E. 2005. Species abundance distributions: pattern or process? Funct Ecol 19:177–181.

Magurran, A. E. and P. A. Henderson. 2003. Explaining the excess of rare species in natural species abundance distributions. Nature 422:714–716.

Makarieva, A. M., V. G. Gorshkov, and B.-L. Li. 2005. Revising the distributive networks models of West, Brown and Enquist (1997) and Banavar, Maritan and Rinaldo (1999): metabolic inequity of living tissues provides clues for the observed allometric scaling rules. J Theor Biol 237:291–301.

Margoliash, E. 1963. Primary structure and evolution of cytochrome c. Proc Natl Acad Sci U S A 50:672–679.

Masel, J. and M. L. Siegal. 2009. Robustness: mechanisms and consequences. Trends Genet 25:395–403.

Mau, B. 1996. Bayesian Phylogenetic Inference via Markov Chain Monte Carlo Methods. Ph.D. thesis University of Wisconsin Madison.

May, R. M. 1975. Patterns of species abundance and diversity. Pages 81–120. Ecology and Evolution of Communities. Harvard University Press, Cambridge.

May, R. M. 1986. The search for patterns in the balance of nature: advances and retreats. Ecology 67:1115–1126.

May, R. M. 1990. Taxonomy as destiny. Nature 347:129–130.

Mayden, R. L. 1997. A hierarchy of species concepts: The denouement of the saga of the species problem. Pages 381–424. Species: The units of biodiversity. Chapman & Hall, New York.

Mayr, E. 1963. Animal species and evolution. Harvard University Press, Cambridge.

Mayr, E. 1982. Speciation and macroevolution. Evolution 36:1119–1132.

Mayr, E. 1991. One Long Argument: Charles Darwin and the Genesis of Modern Evolutionary Thought. Harvard University Press, Cambridge.

McBreen, K. and P. J. Lockhart. 2006. Reconstructing reticulate evolutionary histories of plants. Trends Plant Sci 11:398–404.

McKenzie, A. and M. Steel. 2000. Distributions of cherries for two models of trees. Math Biosci 164:81–92.

McNeill, J., F. R. Barrie, H. M. Burdet, V. Demoulin, D. L. Hawksworth, K. Marhold, D. H. Nicolson, J. Prado, P. C. Silva, J. E. Skog, J. H. Wiersema, and N. J. Turland, eds. 2007. International Code of Botanical Nomenclature (Vienna Code) adopted by the Seventeenth International Botanical Congress Vienna, Austria, July 2005. vol. 146 of *Regnum Vegetabile*. Gantner Verlag KG, Ruggell.

Mendel, J. G. 1865. Versuche über Pflanzen-Hybriden. Verhandlungen des naturforschenden Vereines in Brünn IV:3–47.

Mindell, D. P. and A. Meyer. 2001. Homology evolving. Trends Ecol Evol 16:434–440.

Mitchell, M. 2009. Complexity: A Guided Tour. Oxford University Press, New York.

Montoya, J. M., S. L. Pimm, and R. V. Solé. 2006. Ecological networks and their fragility. Nature 442:259–264.

Montoya, J. M. and R. V. Solé. 2002. Small world patterns in food webs. J Theor Biol 214:405–412.

Montoya, J. M., G. Woodward, M. C. Emmerson, and R. V. Solé. 2009. Press perturbations and indirect effects in real food webs. Ecology 90:2426–2433.

Mooers, A. Ø. and S. B. Heard. 1997. Inferring evolutionary process from the phylogenetic tree shape. Q Rev Biol 72:31–54.

Mooers, A. Ø., R. D. M. Page, A. Purvis, and P. H. Harvey. 1995. Phylogenetic noise leads to unbalanced cladistic trees reconstructions. Syst Biol 44:332–342.

Moore, B. R. 2007. Inferring patterns of diversification. Pages 178–181. Encyclopedia of Science and Technology. McGraw Hill, New York.

Morris, R. J., O. T. Lewis, and H. C. J. Godfray. 2004. Experimental evidence for apparent competition in a tropical forest food web. Nature 428:310–313.

Morris, S. C. 2000. Evolution: bringing molecules into the fold. Cell 100:1–11.

Morrison, D. A. 2005. Networks in phylogenetic analysis: new tools for population biology. Int J Parasitol 35:567–582.

Müller, C. B., I. C. T. Adriaanse, R. Belshaw, and H. C. J. Godfray. 1999. The structure of an aphid-parasitoid community. J Anim Ecol 68:346–370.

Mullis, K. B. 1990. The unusual origin of the polymerase chain reaction. Sci Am 262:56–61, 64–65.

Multitree. 2010. http://multitree.linguistlist.org/.

Murray, C. D. 1926. The Physiological Principle of Minimum Work: I. The Vascular System and the Cost of Blood Volume. Proc Natl Acad Sci U S A 12:207–214.

Nee, S. 2001. Inferring speciation rates from phylogenies. Evolution 55:661–668.

Nee, S., E. C. Holmes, R. M. May, and P. H. Harvey. 1994a. Extinction rates can be estimated from molecular phylogenies. Philos Trans R Soc Lond B Biol Sci 344:77–82.

Nee, S., R. M. May, and P. H. Harvey. 1994b. The reconstructed evolutionary process. Philos Trans R Soc Lond B Biol Sci 344:305–311.

Nei, M. 2005. Selectionism and Neutralism in Molecular Evolution. Mol Biol Evol 22:2318–2342.

Nelson, G. 1979. Cladistic analysis and synthesis: Principles and definitions, with a historical note on Adanson's *Families des Plantes* (1763-1764). Syst Zool 28:1–21.

NemATOL. 2010. http://nematol.unh.edu/.

Newman, M. E. J. 2003. The Structure and Function of Complex Networks. SIAM Review 45:167–256.

Nixon, K. and J. Carpenter. 2000. On the other "phylogenetic systematics". Cladistics 16:298–318.

Nuttall, G. H. 1904. Blood immunity and blood relationship. Cambridge University Press, Cambridge.

Ohta, T. 1992. The nearly neutral theory of molecular evolution. Ann Rev Ecol Syst 23:263–286.

Ohta, T. 1996a. The current significance and standing of neutral and neutral theories. Bioessays 18:673–677; discussion 683.

Ohta, T. 1996b. The neutralist-selectionist debate. Bioessays 18:673–684.

Page, R. D. M. 1998. Molecular evolution: A phylogenetic approach. Blackwell Science, Oxford.

Pagel, M., C. Venditti, and A. Meade. 2006. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. Science 314:119–121.

PANDIT. 2010. http://www.ebi.ac.uk/goldman-srv/pandit/.

Paradis, E. 2008. Asymmetries in phylogenetic diversification and character change can be untangled. Evolution 62:241–247.

Penny, D., B. J. McComish, M. A. Charleston, and M. D. Hendy. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J Mol Evol 53:711–723.

Penny, D. and M. J. Phillips. 2004. The rise of birds and mammals: are microevolutionary processes sufficient for macroevolution? Trends Ecol Evol 19:516–522.

Pfam. 2010. http://pfam.sanger.ac.uk/.

Phylogeny_of_Spiders. 2010. http://research.amnh.org/atol/files/.

PhylomeDB. 2010. http://phylomedb.org/.

Pigolotti, S., A. Flammini, M. Marsili, and A. Maritan. 2005. Species lifetime distribution for simple models of ecologies. Proc Natl Acad Sci U S A 102:15747–15751.

Pimm, S. L., G. J. Russell, J. L. Gittleman, and T. M. Brooks. 1995. The future of biodiversity. Science 269:347–350.

Pinelis, I. 2003. Evolutionary models of phylogenetic trees. Proc Biol Sci 270:1425–1431.

Pons, J., T. G. Barraclough, J. Gomez-Zurita, A. Cardoso, D. P. Duran, S. Hazell, S. Kamoun, W. D. Sumlin, and A. P. Vogler. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst Biol 55:595–609.

Popovic, L. 2004. Asymptotic genealogy of a Critical Branching Process. Ann Appl Probab 14:2120–2148.

Posada, D. and K. A. Crandall. 2001. Intraspecific gene genealogies: trees grafting into networks. Trends Ecol Evol 16:37–45.

Preston, F. W. 1948. The commonness, and rarity, of species. Ecology 29:254–283.

Proulx, S. R., D. E. L. Promislow, and P. C. Phillips. 2005. Network thinking in ecology and evolution. Trends Ecol Evol 20:345–353.

Purvis, A. and P.-M. Agapow. 2002. Phylogeny imbalance: taxonomic level matters. Syst Biol 51:844–854.

Purvis, A. and T. J. Garland. 1993. Polytomies in Comparative Analyses of Continuous Characters. Syst Biol 42:569–575.

Purvis, A. and A. Hector. 2000. Getting the measure of biodiversity. Nature 405:212–219.

Pybus, O. G. and P. H. Harvey. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc Biol Sci 267:2267–2272.

Qiao, B., T. L. Goldberg, G. J. Olsen, and R. M. Weigel. 2006. A computer simulation analysis of the accuracy of partial genome sequencing and restriction fragment analysis in the reconstruction of phylogenetic relationships. Infect Genet Evol 6:323–330.

Ragan, M. A. 2009. Trees and networks before and after Darwin. Biol Direct 4:43.

Ragan, M. A., J. O. McInerney, and J. A. Lake. 2009. The network of life: genome beginnings and evolution. Introduction. Philos Trans R Soc Lond B Biol Sci 364:2169–2175.

Rannala, B. and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol 43:304–311.

Reed, W. J. and B. D. Hughes. 2002. On the size distribution of live genera. J Theor Biol 217:125–135.

Reed, W. J. and B. D. Hughes. 2007. Theoretical size distribution of fossil taxa: analysis of a null model. Theor Biol Med Model 4:12.

Reznick, D. N. and R. E. Ricklefs. 2009. Darwin's bridge between microevolution and macroevolution. Nature 457:837–842.

Ribera, I., T. G. Barraclough, and A. P. Vogler. 2001. The effect of habitat type on speciation rates and range movements in aquatic beetles: inferences from species-level phylogenies. Mol Ecol 10:721–735.

Richter, J. P., ed. 1939. The Literary Works of Leonardo da Vinci. Oxford University Press, London.

Ricklefs, R. E. 2007. Estimating diversification rates from phylogenetic information. Trends Ecol Evol 22:601–610.

Ride, W. D. L., H. G. Cogger, C. Dupuis, O. Kraus, A. Minelli, F. C. Thompson, and P. K. Tubbs, eds. 1999. International Commission on Zoological Nomenclature: International Code of Zoological Nomenclature. 4th ed. The International Trust for Zoological Nomenclature 1999, London.

Rieppel, O. 2005. Monophyly, paraphyly, and natural kinds. Biol Philos 20:465–487.

Rieppel, O. 2006a. The PhyloCode: A critical discussion of its theoretical foundation. Cladistics 22:186–197.

Rieppel, O. 2006b. The taxonomic hierarchy. Systematist 26:5–9.

Rivera, M. C. 2007. Genomic analyses and the origin of the eukaryotes. Chem Biodivers 4:2631–2638.

Rivera, M. C. and J. A. Lake. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431:152–155.

Rodriguez-Iturbe, I. and A. Rinaldo. 1997. Fractal river basins: chance and self-organization. Cambridge University Press, New York.

Rohlf, F. J., W. S. Chang, R. R. Sokal, and J. Kim. 1990. Accuracy of estimated phylogenies: Effects of tree topology and evolutionary model. Evolution 44:1671–1684.

Rokas, A. 2006. Genomics. Genomics and the tree of life. Science 313:1897–1899.

Romanes, G. J. 1895. Darwin and after Darwin, vol 2. Open Court, Chicago.

Rosenzweig, M. L. 1995. Species diversity in space and time. Cambridge University Press, Cambridge.

Roth, C., S. Rastogi, L. Arvestad, K. Dittmar, S. Light, D. Ekman, and D. A. Liberles. 2007. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. J Exp Zool B Mol Dev Evol 308:58–73.

Roux, W. 1878. On the bifurcation of blood vessels. Ph.D. thesis University of Jena Jena.

Rozenfeld, A. F., S. Arnaud-Haond, E. Hernández-García, V. M. Eguíluz, M. A. Matías, E. Serrão, and C. M. Duarte. 2007. Spectrum of genetic diversity and networks of clonal organisms. J R Soc Interface 4:1093–1102.

Rozenfeld, A. F., S. Arnaud-Haond, E. Hernández-García, V. M. Eguíluz, E. A. Serrão, and C. M. Duarte. 2008. Network analysis identifies weak and strong links in a metapopulation system. Proc Natl Acad Sci U S A 105:18824–18829.

Ruan, J., H. Li, Z. Chen, A. Coghlan, L. J. M. Coin, Y. Guo, J.-K. Hériché, Y. Hu, K. Kristiansen, R. Li, T. Liu, A. Moses, J. Qin, S. Vang, A. J. Vilella, A. Ureta-Vidal, L. Bolund, J. Wang, and R. Durbin. 2008. TreeFam: 2008 Update. Nucleic Acids Res 36:D735–D740.

Ryle, A. P., F. Sanger, L. F. Smith, and R. Kitai. 1955. The disulphide bonds of insulin. Biochem J 60:541–556.

Sackin, M. 1972. Good and bad phenograms. Syst Zool 21:225–226.

Sahney, S., M. J. Benton, and P. A. Ferry. 2010. Links between global taxonomic diversity, ecological diversity and the expansion of vertebrates on land. Biol Lett 6:544–547.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425.

Salemi, M. and A.-M. Vandamme, eds. 2003. The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny. Cambridge University Press, Cambridge.

Salisbury, B. A. 1999. Misinformative characters and phylogeny shape. Syst Biol 48:153–169.

Samper, C. 2004. Taxonomy and environmental policy. Philos Trans R Soc Lond B Biol Sci 359:721–728.

Sanderson, M. J. 1996. Reconstructing shifts in diversification rates on phylogenetic trees. Trends Ecol Evol 11:15–20.

Sanderson, M. J., M. J. Donoghue, W. Piel, and T. Eriksson. 1994. TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. Amer Jour Bot 81:183.

Savage, H. M. 1983. The shape of evolution: systematic tree topology. Biol J Linnean Soc 20:225–244.

Savolainen, V., S. B. Heard, M. P. Powell, T. J. Davies, and A. Ø. Mooers. 2002. Is cladogenesis heritable? Syst Biol 51:835–843.

Schlick-Steiner, B. C., F. M. Steiner, B. Seifert, C. Stauffer, E. Christian, and R. H. Crozier. 2010. Integrative taxonomy: a multisource approach to exploring biodiversity. Annu Rev Entomol 55:421–438.

Schluter, D. 1996. Ecological causes of adaptive radiation. Am Nat 148(suppl.):S40–S64.

Schluter, D. 2000. The ecology of adaptive radiation. Oxford University Press, Oxford.

Schwartz, M. W. and D. Simberloff. 2001. Taxon size predicts rates of rarity in vascular plants. Ecol Lett 4:464–469.

Shao, K. and R. R. Sokal. 1990. Tree Balance. Syst Zool 39:266–276.

Simberloff, D., K. L. Hecht, E. D. McCoy, and E. F. Conner. 1981. There have been no statistical tests of cladistic biogeographical hypotheses. Pages 40–63. Vicariance biogeography: a critique. Columbia University Press, New York.

Simon, H. A. 1995. On a class of skew distribution functions. Biometrika 42:425–440.

Simons, A. M. 2002. The continuity of microevolution and macroevolution. J Evol Biol 15:688–701.

Simpson, G. G. 1953. The major features of evolution. Columbia University Press, New York.

Slowinski, J. B. and C. Guyer. 1989. Testing the stochasticity of patterns of organismal diversity: an improved null model. Am Nat 134:907–921.

Sokal, R. R. and C. D. Michener. 1958. A Statistical Method for Evaluating Systematic Relationships. Univ Kansas Sci Bull 38:1409–1438.

Solé, R. V., R. Pastor-Satorras, E. D. Smith, and T. Kepler. 2002. A model of large-scale proteome evolution. Adv Comp Syst 5:43–54.

Soria-Carrasco, V., G. Talavera, J. Igea, and J. Castresana. 2007. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics 23:2954–2956.

Spencer, M., E. Susko, and A. J. Roger. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol 22:1161–1164.

Steel, M. and D. Penny. 2010. Origins of life: Common ancestry put to the test. Nature 465:168–169.

Stich, M. and S. C. Manrubia. 2009. Topological properties of phylogenetic trees in evolutionary models. Eur Phys J B 71:583–592.

Stumpf, M. P. H., W. P. Kelly, T. Thorne, and C. Wiuf. 2007. Evolution at the system level: the natural history of protein interaction networks. Trends Ecol Evol 22:366–373.

Sullivan, J. P., J. G. Lundberg, and M. Hardman. 2006. A phylogenetic analysis of the major groups of catfishes (Teleostei: Siluriformes) using rag1 and rag2 nuclear gene sequences. Mol Phylogenet Evol 41:636–662.

Suthers, P. F., A. Zomorrodi, and C. D. Maranas. 2009. Genome-scale gene/reaction essentiality and synthetic lethality analysis. Mol Syst Biol 5:301.

SYSTERS. 2010. http://systers.molgen.mpg.de/.

Szabó, G., M. Alava, and J. Kertész. 2002. Shortest paths and load scaling in scale-free trees. Phys Rev E 66:026101.

Tamarin, R. H. 1996. Principios de Genética. Editorial Reverté. S.A., Barcelona.

Tamura, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol 9:678–687.

Tamura, K. and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526.

Tateno, Y., M. Nei, and F. Tajima. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J Mol Evol 18:387–404.

Templado, J. 1982. Historia de las teorías evolucionistas. Alhambra, Madrid.

The_Beetle_Tree_of_Life_Project. 2010. http://insects.oeb.harvard.edu/ATOL/.

The_Green_Tree_of_Life. 2010. http://ucjeps.berkeley.edu/TreeofLife/.

The_Mammal_Tree_of_Life. 2010. http://mammaltree.informatics.-sunysb.edu/.

Timetree. 2010. http://www.timetree.org/.

Tokeshi, M. 1993. Species abundance patterns and community structure. Adv Ecol Res 24:112–186.

Tokeshi, M. 1996. Power fraction: a new explanation for species abundance patterns in species-rich assemblages. Oikos 75:543–550.

TreeBASE. 2010. http://www.phylo.org/treebase/home.php.

TreeFam. 2010. http://www.treefam.org/.

Tree_of_Life_Web_Project. 2010. http://tolweb.org/tree/.

Tuffley, C. and M. Steel. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math Biosci 147:63–91.

Unger, R., S. Uliel, and S. Havlin. 2003. Scaling law in sizes of protein sequence families: from super-families to orphan genes. Proteins 51:569–576.

Vázquez, D. P., N. P. Chacoff, and L. Cagnolo. 2009. Evaluating multiple determinants of the structure of plant-animal mutualistic networks. Ecology 90:2039–2046.

Vázquez, D. P. and J. L. Gittleman. 1998. Biodiversity conservation: does phylogeny matter? Curr Biol 8:R379–R381.

Vázquez, D. P., R. Poulin, B. R. Krasnov, and G. Shenbrot. 2005. Species abundance and the distribution of specialization in host-parasite interaction networks. J Anim Ecol 74:946–955.

Veen, F. J. F. V., C. B. Müller, J. K. Pell, and H. C. J. Godfray. 2008. Food web structure of three guilds of natural enemies: predators, parasitoids and pathogens of aphids. J Anim Ecol 77:191–200.

Venditti, C., A. Meade, and M. Pagel. 2010. Phylogenies reveal new interpretation of speciation and the Red Queen. Nature 463:349–352.

Volkov, I., J. R. Banavar, S. P. Hubbell, and A. Maritan. 2003. Neutral theory and relative species abundance in ecology. Nature 424:1035–1037.

von Linné, C. 1758. Systema Naturæ Per Regna Tria Naturæ, Secundum Classes, Ordines, Genera, Species Cum Characteribus, Differentiis, Synonymis, Locis. 10th ed. Laurentii Salvii, Holmiae.

Wagner, A. 2003. How the global structure of protein interaction networks evolves. Proc Biol Sci 270:457–466.

Wagner, A. 2005. Robustness and evolvability in living systems. Princeton University Press, Princeton.

Wagner, A. 2008a. Neutralism and selectionism: a network-based reconciliation. Nat Rev Genet 9:965–974.

Wagner, A. 2008b. Robustness and evolvability: a paradox resolved. Proc Biol Sci 275:91–100.

Wahrmund, U., D. Quandt, and V. Knoop. 2010. The phylogeny of mosses - addressing open issues with a new mitochondrial locus: group I intron cobi420. Mol Phylogenet Evol 54:417–426.

Wake, D. B. and V. T. Vredenburg. 2008. Are we in the midst of the sixth mass extinction? A view from the world of amphibians. Proc Nat Acad Sci USA 105:11466–11473.

Wallace, A. R. 1889. Darwinism : an exposition of the theory of natural selection, with some of its applications. Macmillan, London.

Wang, H.-C., M. Spencer, E. Susko, and A. J. Roger. 2007. Testing for covarion-like evolution in protein sequences. Mol Biol Evol 24:294–305.

Waser, N. M. 2006. Specialization and generalization in plant-pollinator interactions: historical perspective. Pages 3–17. Plant-Pollinator Interactions: From Specialization to Generalization. University of Chicago Press Ltd, London.

Watson, J. D. and F. H. C. Crick. 1953. A Structure for Deoxyribose Nucleic Acid. Nature 171:737–738.

Watts, D. J. and S. H. Strogatz. 1998. Collective dynamics of 'small-world' networks. Nature 393:440–442.

Webb, J. K., B. W. Brook, and R. Shine. 2002. What makes a species vulnerable to extinction? Comparative life-history traits of two sympatric snakes. Ecol Res 17:59–67.

Webvision. 2010. http://webvision.umh.es/webvision/imageswv/-huretina.jpeg.

Weinberg, W. 1908. Über den Nachweis der Vererbung beim Menschen. Jahresh Verein f vaterl Naturk Württem 64:368–382.

Weismann, A. 1892. Das Keimplasma. Eine Theorie der Vererbung. Fischer, Jena.

Wells, W. C. 1818. Two essays: one upon a single vision with two eyes; the other on dew. Archibald Constable and Co. Edinburgh, London.

West, G. B. and J. H. Brown. 2005. The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization. J Exp Biol 208:1575–1592.

West, G. B., J. H. Brown, and B. J. Enquist. 1997. A general model for the origin of allometric scaling laws in biology. Science 276:122–126.

West, G. B., J. H. Brown, and B. J. Enquist. 1999. The fourth dimension of life: fractal geometry and allometric scaling of organisms. Science 284:1677–1679.

Whelan, S., P. I. W. de Bakker, and N. Goldman. 2003. Pandit: a database of protein and associated nucleotide domains with inferred trees. Bioinformatics 19:1556–1563.

Whelan, S., P. I. W. de Bakker, E. Quevillon, N. Rodriguez, and N. Goldman. 2006. PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. Nucleic Acids Res 34:D327–D331.

Whittaker, R. H. 1969. New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. Science 163:150–160.

Williams, C. B. 1964. Patterns in the balance of nature and related problems in quantitative ecology. Academic Press, New York.

Willis, J. C. 1922. Age and area: a study in geographical distribution and origin of species. Cambridge University Press, Cambridge.

Winchester, W. 2001. The Map that Changed the World: William Smith and the Birth of Modern Geology. Harper Collins, New York.

Woese, C. R. and G. E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74:5088–5090.

Woese, C. R., O. Kandler, and M. L. Wheelis. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 87:4576–4579.

Woodward, G. 2008. Biodiversity, ecosystem functioning and food webs in freshwaters: assembling the jigsaw puzzle. Freshwater Biol 54:2171–2187.

Woodward, G., B. Ebenman, M. Emmerson, J. M. Montoya, J. M. Olesen, A. Valido, and P. H. Warren. 2005a. Body size in ecological networks. Trends Ecol Evol 20:402–409.

Woodward, G., D. Speirs, and A. G. Hildrew. 2005b. Quantification and resolution of complex, size-structured food web. Adv Ecol Res 36:85–135.

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

Wright, S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proceedings of the Sixth International Congress on Genetics 1:356–366.

Yamada, T. and P. Bork. 2009. Evolution of biomolecular networks: lessons from metabolic and protein interactions. Nat Rev Mol Cell Biol 10:791–803.

Yang, S. and P. E. Bourne. 2009. The evolutionary history of protein domains viewed by species phylogeny. PLoS One 4:e8378.

Yi, Z., W. Song, J. C. Clamp, Z. Chen, S. Gao, and Q. Zhang. 2009. Reconsideration of systematic relationships within the order Euplotida (Protista, Ciliophora) using new sequences of the gene coding for small-subunit rRNA and testing the use of combined data sets to construct phylogenies of the Diophrys-complex. Mol Phylogenet Evol 50:599–607.

Yule, G. U. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. Philos Trans R Soc Lond A 213:21–87.

Zeh, D. W., J. A. Zeh, and Y. Ishida. 2009. Transposable elements and an epigenetic basis for punctuated equilibria. Bioessays 31:715–726.

Zhang, J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol 18:292–298.

Zhang, J. 2009. Allometric Scaling of Weighted Food Webs. vol. 5 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. Pages 1441–1450. Springer Verlag, Berlin Heidelberg.

Zhang, J. and L. Guo. 2010. Scaling behaviors of weighted food webs as energy transportation networks. J Theor Biol 264:760–770.

Zhang, Z., S. Zhou, L. Chen, J. Guan, L. Fang, and Y. Zhang. 2007. Recursive weighted treelike networks. Eur Phys J B 59:99–107.

Zhou, T., D. A. Drummond, and C. O. Wilke. 2008. Contact density affects protein evolutionary rate from bacteria to animals. J Mol Evol 66:395–404.

Zhou, Y., H. Brinkmann, N. Rodrigue, N. Lartillot, and H. Philippe. 2010. A Dirichlet process covarion mixture model and its assessments using posterior predictive discrepancy tests. Mol Biol Evol 27:371–384.

Zhou, Y., N. Rodrigue, N. Lartillot, and H. Philippe. 2007. Evaluation of the models handling heterotachy in phylogenetic inference. BMC Evol Biol 7:206.

Zima, J. and I. Horaček. 1978. Factors of diversification of mammalian taxa. Proc. Symp. Natur. Select. Loblice. CSAV, Praha.

Zink, R. M. and J. B. Slowinski. 1995. Evidence from molecular systematics for decreased avian diversification in the pleistocene Epoch. Proc Natl Acad Sci U S A 92:5832–5835.

Zuckerkandl, E. and L. Pauling. 1962. Molecular disease, evolution, and genetic heterogeneity. Pages 189–225. Horizons in Biochemistry. Academic Press, New York.

Zuckerkandl, E. and L. Pauling. 1965. Evolutionary divergence and convergence in proteins. Pages 97–166. Evolving Genes and Proteins. Academic Press, New York.

References of the compiled phylogenies

Albach, D. C., P. Schönswetter, and A. Tribsch. 2006. Comparative phylogeography of the Veronica alpina complex in Europe and North America. Mol Ecol 15:3269–3286.

Andreasen, K. and B. Bremer. 2000. Combined phylogenetic analysis in the Rubiaceae-Ixoroideae: morphology, nuclear and chloroplast DNA data. Am J Bot 87:1731–1748.

Benz, B. W., M. B. Robbins, and A. T. Peterson. 2006. Evolutionary history of woodpeckers and allies (Aves: Picidae): placing key taxa on the phylogenetic tree. Mol Phylogenet Evol 40:389–399.

Beszteri, B., E. Acs, and L. K. Medlin. 2005. Ribosomal DNA sequence variation among sympatric strains of the Cyclotella meneghiniana complex (Bacillariophyceae) reveals cryptic diversity. Protist 156:317–333.

Brindefalk, B., J. Viklund, D. Larsson, M. Thollesson, and S. G. E. Andersson. 2007. Origin and evolution of the mitochondrial aminoacyl-tRNA synthetases. Mol Biol Evol 24:743–756.

Choi, Y.-J., S.-B. Hong, and H.-D. Shin. 2006. Genetic diversity within the Albugo candida complex (Peronosporales, Oomycota) inferred from phylogenetic analysis of ITS rDNA and COX2 mtDNA sequences. Mol Phylogenet Evol 40:400–409.

Cupolillo, E., L. R. Brahim, C. B. Toaldo, M. P. de Oliveira-Neto, M. E. F. de Brito, A. Falqueto, M. de Farias Naiff, and G. Grimaldi. 2003. Genetic polymorphism and molecular epidemiology of Leishmania (Viannia) braziliensis from different hosts and geographic areas in Brazil. J Clin Microbiol 41:3126–3132.

de Casas, R. R., G. Besnard, P. Schönswetter, L. Balaguer, and P. Vargas. 2006. Extensive gene flow blurs phylogeographic but not phylogenetic signal in Olea europaea L. Theor Appl Genet 113:575–583.

Devitt, T. J. 2006. Phylogeography of the Western Lyresnake (Trimorphodon biscutatus): testing aridland biogeographical hypotheses across the Nearctic-Neotropical transition. Mol Ecol 15:4387–4407.

Dighe, A. S., K. Jangid, J. M. González, V. J. Pidiyar, M. S. Patole, D. R. Ranade, and Y. S. Shouche. 2004. Comparison of 16S rRNA gene sequences of genus Methanobrevibacter. BMC Microbiol 4:20.

Dohrmann, M., O. Voigt, D. Erpenbeck, and G. Wörheide. 2006. Non-monophyly of most supraspecific taxa of calcareous sponges (Porifera, Calcarea) revealed by increased taxon sampling and partitioned Bayesian analysis of ribosomal DNA. Mol Phylogenet Evol 40:830–843.

Driscoll, D. A. and C. M. Hardy. 2005. Dispersal and phylogeography of the agamid lizard Amphibolurus nobbi in fragmented and continuous habitat. Mol Ecol 14:1613–1629.

Duda, T. F. and A. J. Kohn. 2005. Species-level phylogeography and evolutionary history of the hyperdiverse marine gastropod genus Conus. Mol Phylogenet Evol 34:257–272.

Ehling-Schulz, M., B. Svensson, M.-H. Guinebretiere, T. Lindbäck, M. Andersson, A. Schulz, M. Fricker, A. Christiansson, P. E. Granum, E. Märtlbauer, C. Nguyen-The, M. Salkinoja-Salonen, and S. Scherer. 2005. Emetic toxin formation of Bacillus cereus is restricted to a single evolutionary lineage of closely related strains. Microbiology 151:183–197.

Ellison, N. W., A. Liston, J. J. Steiner, W. M. Williams, and N. L. Taylor. 2006. Molecular phylogenetics of the clover genus (Trifolium–Leguminosae). Mol Phylogenet Evol 39:688–705.

Endress, P. K. and J. A. Doyle. 2007. Floral phyllotaxis in basal angiosperms: development and evolution. Curr Opin Plant Biol 10:52–57.

Fitzpatrick, D. A., M. E. Logue, J. E. Stajich, and G. Butler. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol Biol 6:99.

Fuchs, J., C. Cruaud, A. Couloux, and E. Pasquet. 2007. Complex biogeographic history of the cuckoo-shrikes and allies (Passeriformes: Campephagidae) revealed by mitochondrial and nuclear sequence data. Mol Phylogenet Evol 44:138–153.

Fuchs, J., E. Pasquet, A. Couloux, J. Fjeldså, and R. C. K. Bowie. 2009. A new Indo-Malayan member of the Stenostiridae (Aves: Passeriformes) revealed by multilocus sequence data: biogeographical implications for a morphologically diverse clade of flycatchers. Mol Phylogenet Evol 53:384–393.

Fuchs, J., J.-M. Pons, S. M. Goodman, V. Bretagnolle, M. Melo, R. C. K. Bowie, D. Currie, R. Safford, M. Z. Virani, S. Thomsett, A. Hija, C. Cruaud, and E. Pasquet. 2008. Tracing the colonization history

of the Indian Ocean scops-owls (Strigiformes: Otus) with further insight into the spatio-temporal origin of the Malagasy avifauna. BMC Evol Biol 8:197.

Fulton, T. L. and C. Strobeck. 2006. Molecular phylogeny of the Arctoidea (Carnivora): effect of missing data on supertree and supermatrix analyses of multiple gene data sets. Mol Phylogenet Evol 41:165–181.

Gamage, D. T., M. P. de Silva, N. Inomata, T. Yamazaki, and A. E. Szmidt. 2006. Comprehensive molecular phylogeny of the subfamily Dipterocarpoideae (Dipterocarpaceae) based on chloroplast DNA sequences. Genes Genet Syst 81:1–12.

García, D., A. M. Stchigel, J. Cano, M. Calduch, D. L. Hawksworth, and J. Guarro. 2006. Molecular phylogeny of Coniochaetales. Mycol Res 110:1271–1289.

Garcia, J. L., B. K. Patel, and B. Ollivier. 2000. Taxonomic, phylogenetic, and ecological diversity of methanogenic Archaea. Anaerobe 6:205–226.

Gast, R. J. 2006. Molecular phylogeny of a potentially parasitic dinoflagellate isolated from the solitary radiolarian, Thalassicolla nucleata. J Eukaryot Microbiol 53:43–45.

Gaubert, P. and P. Cordeiro-Estrela. 2006. Phylogenetic systematics and tempo of evolution of the Viverrinae (Mammalia, Carnivora, Viverridae) within feliformians: implications for faunal exchanges between Asia and Africa. Mol Phylogenet Evol 41:266–278.

Gottschling, M., A. Köhler, E. Stockfleth, and I. Nindl. 2007. Phylogenetic analysis of beta-papillomaviruses as inferred from nucleotide and amino acid sequence data. Mol Phylogenet Evol 42:213–222.

Gray, R. D. and Q. D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426:435–439.

Gray, R. D. and F. M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. Nature 405:1052–1055.

Habayeb, M. S., S. K. Ekengren, and D. Hultmark. 2006. Nora virus, a persistent virus in Drosophila, defines a new picorna-like virus family. J Gen Virol 87:3045–3051.

Hahn, M. W., M. Pöckl, and Q. L. Wu. 2005. Low intraspecific diversity in a polynucleobacter subcluster population numerically dominating bacterioplankton of a freshwater pond. Appl Environ Microbiol 71:4539–4547.

Hahn, W. J. 2002. A molecular phylogenetic study of the Palmae (Arecaceae) based on atpB, rbcL, and 18S nrDNA sequences. Syst Biol 51:92–112.

Heilveil, J. S. and S. H. Berlocher. 2006. Phylogeography of postglacial range expansion in Nigronia serricornis Say (Megaloptera: Corydalidae). Mol Ecol 15:1627–1641.

Hommais, F., S. Pereira, C. Acquaviva, P. Escobar-Páramo, and E. Denamur. 2005. Single-nucleotide polymorphism phylotyping of Escherichia coli. Appl Environ Microbiol 71:4784–4792.

Huang, L.-N., S. Zhu, H. Zhou, and L.-H. Qu. 2005. Molecular phylogenetic diversity of bacteria associated with the leachate of a closed municipal solid waste landfill. FEMS Microbiol Lett 242:297–303.

Huang, S., Y. C. Chiang, B. A. Schaal, C. H. Chou, and T. Y. Chiang. 2001. Organelle DNA phylogeography of Cycas taitungensis, a relict species in Taiwan. Mol Ecol 10:2669–2681.

Humbert, J. F., D. Duris-Latour, B. L. Berre, H. Giraudet, and M. J. Salençon. 2005. Genetic diversity in Microcystis populations of a French storage reservoir assessed by sequencing of the 16S-23S rRNA intergenic spacer. Microb Ecol 49:308–314.

Hyvönen, J., S. Koskinen, G. L. S. Merrill, T. A. Hedderson, and S. Stenroos. 2004. Phylogeny of the Polytrichales (Bryophyta) based on simultaneous analysis of molecular and morphological data. Mol Phylogenet Evol 31:915–928.

Jensen, L. H., H. Enghoff, J. Frydenberg, and E. D. Parker. 2002. Genetic diversity and the phylogeography of parthenogenesis: comparing bisexual and thelytokous populations of Nemasoma varicorne (Diplopoda: Nemasomatidae) in Denmark. Hereditas 136:184–194.

Kawamoto, Y., T. Shotake, K. Nozawa, S. Kawamoto, K. ichiro Tomari, S. Kawai, K. Shirai, Y. Morimitsu, N. Takagi, H. Akaza, H. Fujii, K. Hagihara, K. Aizawa, S. Akachi, T. Oi, and S. Hayaishi. 2007. Postglacial population expansion of Japanese macaques (Macaca fuscata) inferred from mitochondrial DNA phylogeography. Primates 48:27–40.

Ko, K. S., H. K. Lee, M.-Y. Park, and Y.-H. Kook. 2003. Mosaic structure of pathogenicity islands in Legionella pneumophila. J Mol Evol 57:63–72.

Lavoué, S., M. Miya, K. Saitoh, N. B. Ishiguro, and M. Nishida. 2007. Phylogenetic relationships among anchovies, sardines, herrings and their relatives (Clupeiformes), inferred from whole mitogenome sequences. Mol Phylogenet Evol 43:1096–1105.

Le, M., C. J. Raxworthy, W. P. McCord, and L. Mertz. 2006. A molecular phylogeny of tortoises (Testudines: Testudinidae) based on mitochondrial and nuclear genes. Mol Phylogenet Evol 40:517–531.

Lefébure, T., C. J. Douady, M. Gouy, P. Trontelj, J. Briolay, and J. Gibert. 2006. Phylogeography of a subterranean amphipod reveals cryptic diversity and dynamic evolution in extreme environments. Mol Ecol 15:1797–1806.

Li, L., W. Song, A. Warren, Y. Wang, H. Ma, X. Hu, and Z. Chen. 2006. Phylogenetic position of the marine ciliate, Cardiostomatella vermiforme (Kahl, 1928) Corliss, 1960 inferred from the complete SSrRNA gene sequence, with establishment of a new order Loxocephalida n. ord. (Ciliophora, Oligohymenophorea). Eur J Protistol 42:107–114.

Lin, M., D. A. Payne, and J. R. Schwarz. 2003. Intraspecific diversity of Vibrio vulnificus in Galveston Bay water and oysters as determined by randomly amplified polymorphic DNA PCR. Appl Environ Microbiol 69:3170–3175.

Mallatt, J. and G. Giribet. 2006. Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. Mol Phylogenet Evol 40:772–794.

Mann, N. I., F. K. Barker, J. A. Graves, K. A. Dingess-Mann, and P. J. B. Slater. 2006. Molecular data delineate four genera of "Thryothorus" wrens. Mol Phylogenet Evol 40:750–759.

Maraun, M., M. Heethoff, K. Schneider, S. Scheu, G. Weigmann, J. Cianciolo, R. H. Thomas, and R. A. Norton. 2004. Molecular phylogeny of oribatid mites (Oribatida, Acari): evidence for multiple radiations of parthenogenetic lineages. Exp Appl Acarol 33:183–201.

Marimon, R., J. Gené, J. Cano, L. Trilles, M. D. S. Lazéra, and J. Guarro. 2006. Molecular phylogeny of Sporothrix schenckii. J Clin Microbiol 44:3251–3256.

Marmi, J., F. López-Giráldez, D. W. Macdonald, F. Calafell, E. Zholnerovskaya, and X. Domingo-Roura. 2006. Mitochondrial DNA reveals a strong phylogeographic structure in the badger across Eurasia. Mol Ecol 15:1007–1020.

Martínez-Solano, I., J. Teixeira, D. Buckley, and M. García-París. 2006. Mitochondrial DNA phylogeography of Lissotriton boscai

(Caudata, Salamandridae): evidence for old, multiple refugia in an Iberian endemic. Mol Ecol 15:3375–3388.

Michitaka, K., Y. Tanaka, N. Horiike, T. N. Duong, Y. Chen, K. Matsuura, Y. Hiasa, M. Mizokami, and M. Onji. 2006. Tracing the history of hepatitis B virus genotype D in western Japan. J Med Virol 78:44–52.

Miller, C. R., L. P. Waits, and P. Joyce. 2006. Phylogeography and mitochondrial diversity of extirpated brown bear (Ursus arctos) populations in the contiguous United States and Mexico. Mol Ecol 15:4477–4485.

Monis, P. T., R. H. Andrews, G. Mayrhofer, and P. L. Ey. 2003. Genetic diversity within the morphological species Giardia intestinalis and its relationship to host origin. Infect Genet Evol 3:29–38.

Moreira, D., S. von der Heyden, D. Bass, P. López-García, E. Chao, and T. Cavalier-Smith. 2007. Global eukaryote phylogeny: Combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata. Mol Phylogenet Evol 44:255–266.

Moyle, R. G. and B. D. Marks. 2006. Phylogenetic relationships of the bulbuls (Aves: Pycnonotidae) based on mitochondrial and nuclear DNA sequence data. Mol Phylogenet Evol 40:687–695.

Ohlson, J. I., R. O. Prum, and P. G. P. Ericson. 2007. A molecular phylogeny of the cotingas (Aves: Cotingidae). Mol Phylogenet Evol 42:25–37.

Ozeki, M., Y. Isagi, H. Tsubota, P. Jacklyn, and D. M. J. S. Bowman. 2007. Phylogeography of an Australian termite, Amitermes laurensis (Isoptera, Termitidae), with special reference to the variety of mound shapes. Mol Phylogenet Evol 42:236–247.

Perk, S., C. Banet-Noach, E. Shihmanter, S. Pokamunski, M. Pirak, M. Lipkind, and A. Panshin. 2006. Genetic characterization of the

H9N2 influenza viruses circulated in the poultry population in Israel. Comp Immunol Microbiol Infect Dis 29:207–223.

Perneel, M., J. T. Tambong, A. Adiobo, C. Floren, F. Saborío, A. Lévesque, and M. Höfte. 2006. Intraspecific variability of Pythium myriotylum isolated from cocoyam and other host crops. Mycol Res 110:583–593.

Pyron, R. A. and F. T. Burbrink. 2009. Lineage diversification in a widespread species: roles for niche divergence and conservatism in the common kingsnake, Lampropeltis getula. Mol Ecol 18:3443–3457.

Rexová, K., Y. Bastin, and D. Frynta. 2006. Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data. Naturwissenschaften 93:189–194.

Robalo, J. I., V. C. Almada, A. Levy, and I. Doadrio. 2007. Re-examination and phylogeny of the genus Chondrostoma based on mitochondrial and nuclear data and the definition of 5 new genera. Mol Phylogenet Evol 42:362–372.

Roberts, T. E. 2006. History, ocean channels, and distance determine phylogeographic patterns in three widespread Philippine fruit bats (Pteropodidae). Mol Ecol 15:2183–2199.

Rowe, K. C., E. J. Heske, and K. N. Paige. 2006. Comparative phylogeography of eastern chipmunks and white-footed mice in relation to the individualistic nature of species. Mol Ecol 15:4003–4020.

Ruzzante, D. E., S. J. Walde, V. E. Cussac, M. L. Dalebout, J. Seibert, S. Ortubay, and E. Habit. 2006. Phylogeography of the Percichthyidae (Pisces) in Patagonia: roles of orogeny, glaciation, and volcanism. Mol Ecol 15:2949–2968.

Sagegami-Oba, R., Y. Oba, and H. Ohira. 2007. Phylogenetic relationships of click beetles (Coleoptera: Elateridae) inferred from 28S

ribosomal DNA: insights into the evolution of bioluminescence in Elateridae. Mol Phylogenet Evol 42:410–421.

Saldarriaga, J. F., M. L. McEwan, N. M. Fast, F. J. R. Taylor, and P. J. Keeling. 2003. Multiple protein phylogenies show that Oxyrrhis marina and Perkinsus marinus are early branches of the dinoflagellate lineage. Int J Syst Evol Microbiol 53:355–365.

Scott, J. B. and S. Chakraborty. 2006. Multilocus sequence analysis of Fusarium pseudograminearum reveals a single phylogenetic species. Mycol Res 110:1413–1425.

Sjölin, E., C. Erséus, and M. Källersjö. 2005. Phylogeny of Tubificidae (Annelida, Clitellata) based on mitochondrial and nuclear sequence data. Mol Phylogenet Evol 35:431–441.

Sogstad, M. K. R., E. A. Høiby, and D. A. Caugant. 2006. Molecular characterization of non-penicillin-susceptible Streptococcus pneumoniae in Norway. J Clin Microbiol 44:3225–3230.

Sørensen, M. V. and G. Giribet. 2006. A modern approach to rotiferan phylogeny: combining morphological and molecular data. Mol Phylogenet Evol 40:585–608.

Spinks, P. Q. and H. B. Shaffer. 2005. Range-wide molecular analysis of the western pond turtle (Emys marmorata): cryptic variation, isolation by distance, and their conservation implications. Mol Ecol 14:2047–2064.

Stchigel, A. M., J. Cano, A. N. Miller, M. Calduch, and J. Guarro. 2006. Corylomyces: a new genus of Sordariales from plant debris in France. Mycol Res 110:1361–1368.

Sullivan, J. P., J. G. Lundberg, and M. Hardman. 2006. A phylogenetic analysis of the major groups of catfishes (Teleostei: Siluriformes) using rag1 and rag2 nuclear gene sequences. Mol Phylogenet Evol 41:636–662.

Thangadurai, R., S. L. Hoti, N. P. Kumar, and P. K. Das. 2006. Phylogeography of human lymphatic filarial parasite, Wuchereria bancrofti in India. Acta Trop 98:297–304.

Ursenbacher, S., M. Carlsson, V. Helfer, H. Tegelström, and L. Fumagalli. 2006. Phylogeography and Pleistocene refugia of the adder (Vipera berus) as inferred from mitochondrial DNA sequence data. Mol Ecol 15:3425–3437.

van Ee, B. W., N. Jelinski, P. E. Berry, and A. L. Hipp. 2006. Phylogeny and biogeography of Croton alabamensis (Euphorbiaceae), a rare shrub from Texas and Alabama, using DNA sequence and AFLP data. Mol Ecol 15:2735–2751.

Vancanneyt, M., G. Huys, K. Lefebvre, V. Vankerckhoven, H. Goossens, and J. Swings. 2006. Intraspecific genotypic characterization of Lactobacillus rhamnosus strains intended for probiotic use and isolates of human origin. Appl Environ Microbiol 72:5376–5383.

Verovnik, R., B. Sket, and P. Trontelj. 2004. Phylogeography of subterranean and surface populations of water lice Asellus aquaticus (Crustacea: Isopoda). Mol Ecol 13:1519–1532.

Wahrmund, U., D. Quandt, and V. Knoop. 2010. The phylogeny of mosses - addressing open issues with a new mitochondrial locus: group I intron cobi420. Mol Phylogenet Evol 54:417–426.

Wang, Z., M. Binder, C. L. Schoch, P. R. Johnston, J. W. Spatafora, and D. S. Hibbett. 2006. Evolution of helotialean fungi (Leotiomycetes, Pezizomycotina): a nuclear rDNA phylogeny. Mol Phylogenet Evol 41:295–312.

Ward, T. J., L. Gorski, M. K. Borucki, R. E. Mandrell, J. Hutchins, and K. Pupedis. 2004. Intraspecific phylogeny and lineage group identification based on the prfA virulence gene cluster of Listeria monocytogenes. J Bacteriol 186:4994–5002.

Whipps, C. M. and M. L. Kent. 2006. Phylogeography of the cosmopolitan marine parasite Kudoa thyrsites (Myxozoa: Myxosporea). J Eukaryot Microbiol 53:364–373.

Wilson, N. G., M. Schrödl, and K. M. Halanych. 2009. Ocean barriers and glaciation: evidence for explosive radiation of mitochondrial lineages in the Antarctic sea slug Doris kerguelenensis (Mollusca, Nudibranchia). Mol Ecol .

Wright, A.-D. G. 2006. Phylogenetic relationships within the order Halobacteriales inferred from 16S rRNA gene sequences. Int J Syst Evol Microbiol 56:1223–1227.

Yi, Z., W. Song, J. C. Clamp, Z. Chen, S. Gao, and Q. Zhang. 2009. Reconsideration of systematic relationships within the order Euplotida (Protista, Ciliophora) using new sequences of the gene coding for small-subunit rRNA and testing the use of combined data sets to construct phylogenies of the Diophrys-complex. Mol Phylogenet Evol 50:599–607.

Zanatta, D. T. and R. W. Murphy. 2006. Evolution of active host-attraction strategies in the freshwater mussel tribe Lampsilini (Bivalvia: Unionidae). Mol Phylogenet Evol 41:195–208.

Zhang, C., M. P. Mammen, P. Chinnawirotpisan, C. Klungthong, P. Rodpradit, A. Nisalak, D. W. Vaughn, S. Nimmannitya, S. Kalayanarooj, and E. C. Holmes. 2006a. Structure and age of genetic diversity of dengue virus type 2 in Thailand. J Gen Virol 87:873–883.

Zhang, W.-J., J. Yang, Y.-H. Yu, S.-W. Shu, and Y.-F. Shen. 2006b. Population genetic structure of Carchesium polypinum (Ciliophora: Peritrichia) in four Chinese lakes inferred from ISSR fingerprinting: high diversity but low differentiation. J Eukaryot Microbiol 53:358–363.

Zhang, Z., T. Kudo, Y. Nakajima, and Y. Wang. 2001. Clarification of the relationship between the members of the family Thermomonosporaceae on the basis of 16S rDNA, 16S-23S rRNA internal transcribed spacer and 23S rDNA sequences and chemotaxonomic analyses. Int J Syst Evol Microbiol 51:373–383.

Zink, R. M., S. V. Drovetski, and S. Rohwer. 2006. Selective neutrality of mitochondrial ND2 sequences, phylogeography and species limits in Sitta europaea. Mol Phylogenet Evol 40:679–686.

Zorrilla, I., M. A. Moriñigo, D. Castro, M. C. Balebona, and J. J. Borrego. 2003. Intraspecific characterization of Vibrio alginolyticus isolates recovered from cultured fish in Spain. J Appl Microbiol 95:1106–1116.

Zuccon, D., A. Cibois, E. Pasquet, and P. G. P. Ericson. 2006. Nuclear and mitochondrial sequence data reveal the major lineages of starlings, mynas and related taxa. Mol Phylogenet Evol 41:333–344.

"Desde la ventana se veía la ladera en la que crecían los cuerpos retorcidos de los manzanos..." — La insoportable levedad del ser (Milan Kundera, 1984)