1

The Tree of Life (I)

1.1 Introduction

This chapter provides an overview of life and draws near some important questions: When did life on earth begin? What is life? How is it organized? When did multicellular organisms appear and why? How many species exist on Earth? Notions of biology related to the emergence and classification of life are discussed in connection with different strategies on organism formation. Some ultrastructural images (electron microscopy) are presented as examples for reference. The lower and upper physical dimensions of eukaryotic and prokaryotic organisms are explored in detail. The same exploration is made for viruses that interact within the kingdoms of life (Animalia, Plantae, Fungi, Protista, [Archaea and Bacteria] or Monera). Moreover, a discussion closely debates the reference system and the requirements for life; with special considerations for the "spark of metabolism." Next, an introduction is made on some concrete topics, namely: The origins of eukaryotic cells, the endosymbiosis theory, the origins of organelles, the notion of reductive evolution, and the importance of horizontal gene transfer (HGT). Toward the end of the chapter, the main hypotheses regarding the origin of eukaryotic multicellularity are explored using the behavior observed in current species.

1.2 Emergence of Life

The Earth is believed to be ~4.5 billion years old. Geological evidence shows that liquid water, continental crust, and a rudimentary atmosphere existed on Earth just 100 million years later (4.4 billion years ago) [1]. The planetary atmosphere consisted of water vapor, carbon dioxide, methane, and ammonium [2]. It is unknown exactly when or how life began on Earth [3]. It is considered that life began on the early Earth soon after conditions became favorable for a chain of consecutive, yet undetermined chemical reactions [4]. The field of

Algorithms in Bioinformatics: Theory and Implementation, First Edition. Paul A. Gagniuc.

Companion website: www.wiley.com/go/gagniuc/algorithmsinbioinformatics

prebiotic chemistry tries to explain how organic compounds formed in the absence of biology and how these simple molecules self-assembled to ignite life on Earth and possibly on other planets. The oldest fossils of single-celled organisms date around 3.5 billion years ago [5, 6]. Nonetheless, only organisms with a dense biological structure would have resisted the intense metamorphism experienced by crustal rocks for more than 3.5 billion years. In turn, a dense biological structure may indicate high organism complexity. Thus, the earliest known microfossils could actually indicate the presence of structurally complex unicellular organisms [7]. It stands to reason that those "first" organisms must have required a long time to develop their complexity. Evidence for life on Earth before 3.8 billion years ago has been proposed in the past [7]. Preserved carbon, potentially of biogenic nature, pushes the origin of life on Earth to 4.1 billion years [8]. This indicates that life may have occurred fairly quickly after the formation of the planet ($4.5 - 4.1 = 0.4$ or $4.5 - 3.8 = 0.7$). That is, 400 million to 700 million years after the formation of the planet. Moreover, the observation has important implications for our beliefs about how fast life ignites on other planets with similar conditions. In the next important event, life brings chemical modifications to the planetary atmosphere. An oxygen-containing atmosphere and evidence of cyanobacteria and photosynthesis date around 2.4–2.2 billion years ago [9, 10]. Large colonial organisms with coordinated growth in oxygenated environments have been found as far as 2.1 billion years ago [11]. Life made a gradual step toward eukaryotic unicellular organisms ~2 billion years ago. Eukaryotes divide into three main groups around 1.5 billion years ago, namely in the unicellular ancestors of modern plants, fungi, and animals [12, 13]. The appearance of multicellular eukaryotes is an unclear period. During evolution, gain or loss of multicellularity often occurred until a stable multicellular state was reached [14]. Knowledge of the complexity and size of current single-celled eukaryotic organisms calls into question many more complex fossils. It is difficult to investigate whether some macroscopic-sized fossils indicate multicellular or unicellular macroscopic organisms. Some certainty appears in the fossil record with the rise of bilaterians (bilateral symmetry in organisms) over 550–600 million years ago [15]. Nevertheless, even in the case of these fossils, some uncertainty overshadows interpretation. The macroscopic dimensions and the observed bilateral symmetry still cannot indicate with certainty the multicellular nature of these extinct organisms. Again, many of these fossils can be interpreted as multicellular organisms or as unicellular organisms (e.g. giant protists) [16]. More clear evidence suggests that multicellular organisms may have been present around 635 million years ago [17]. Recent molecular clock analyses estimate that animals started to evolve ~650 million years ago [18]. Bilaterian metazoans (animals with bilateral symmetry) first appeared around 600 million years [12]. Moreover, the evolutionary origins of the blood vascular system date around the same period [19]. Later, bilaterians split into the protostomes and deuterostomes [12].

Protostomes give rise to bring about all the arthropods (e.g. insects, spiders, crabs, shrimp, and so on). Deuterostomes eventually give rise to all vertebrates [12]. Perhaps the most important leap made in the evolution of life was the appearance of motility in multicellular organisms. Fossilized trails of bilaterian animals suggest that eukaryotic multicellular organisms have acquired motility around 551–539 million years ago [20]. A "few" million years later, the first true vertebrate with a backbone appears in the fossil record (545–490 million years ago) [21]. Moreover, fossil evidence shows that animals were exploring the land for the first time around 500 million years ago (544–457) [22]. Plants begin colonizing the land around the same time (470 million years ago) [22]. The first four-legged animals (tetrapods) explored the land 385–359 million years ago and gave rise to all amphibians, reptiles, birds, and mammals [22, 23]. The oldest fossilized tree also dates from this period [22]. Important diversifications in eukaryotic species appear after this period, both on land and in water. The first mammal-like forms appear in the fossil record around 225 million years ago [24]. Much closer periods bring many wonders. For instance, the largest eukaryotes in Earth's history have been observed around 100 million years ago, namely the cretaceous dinosaurs (e.g. *Argentinosaurus*; length: 22–35 m; estimated mass: 50 000–100 000 kg) [25]. The last extinction event (Cretaceous–Tertiary extinction), which occurred over 66–65 million years ago, allowed for a relaxed evolution of mammals [26]. Our own story begins of course at the origin, of life. However, more distinguished developments start with the origins of the first primates around 55 million years ago [27]. Note that the timeline of past events is detailed in the literature and here only some general points were reached.

1.2.1 Timeline Disagreements

Microfossils (the imprint left by an organism in stone), stromatolites (layered rocks derived from photosynthetic cyanobacteria remains sedimented over time), sedimentary carbon isotope ratios or molecular fossils derived from cellular and membrane lipids ("biomarkers"), are used for estimations of the origin and diversification of life in the distant past [28]. Data expressed in billions and hundreds of millions of years are particularly subjective and can lead to variations in the literature up to plus or minus half a billion years. These issues are known and must be taken at face value. While timeline estimates may vary, the order of events is particularly objective. Note that timeline disagreements in the paleontology literature rather indicate that evolution has no milestones but trial periods that overlap; some trials more successful and others that we will probably never know about. Nevertheless, the closer the events get to the present, the more reliable the numbers become. Although relative, timeline estimations in paleontology represent a reliable reference system for important past events on our planet.

1.3 Classifications and Mechanisms

Life on Earth was classified by us into three major domains, namely Bacteria, Archaea, and Eukarya (Figure 1.1) [29]. Bacteria (Greek – bakterion, "small stick") and Archaea (Greek – Arkhē, "origin") are prokaryotes (Greek – pro, "before"; karyon, "kernel" or "core"). Prokaryotes are single-cell organisms (unicellular) without a "core," namely without a nucleus, and are considered similar or close in sophistication to the first living organisms on Earth. Eukarya (Greek – eu, "well" or "true") includes all unicellular and multicellular organisms with cells that contain nuclei, and it refers to animals (including us), plants, fungi, and single-celled protists.

All contemporary forms of life store information in DNA molecules. DNA molecules are polymers consisting of four types of organic molecules linked together by phosphate groups, namely: adenine (A), thymine (T), cytosine (C), and guanine (G). Indirectly, all cellular processes are orchestrated by the information contained in the DNA molecule. Cellular processes store and use energy in the form of discrete packets (adenosine triphosphate molecules or ATP). In prokaryotes, DNA shows a double-stranded circular (usually) form

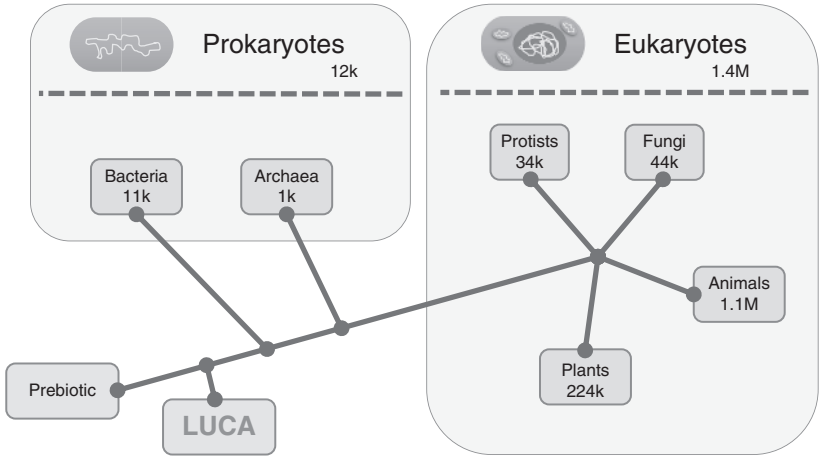

Figure 1.1 The tree of life – basic diagram. The prebiotic period shown on the bottom-left represents the formation of primordial chemical molecules necessary for the ignition of life. Next, the diagram indicates the appearance of LUCA (last universal common ancestor), the first "rudimentary" form of life. The first prokaryotes appear later based on the evolution of LUCA, namely bacteria and archaea. Eukaryotes appear next in the evolutionary chain. Eukaryotes divide the tree of life into four other main subdivisions (eukaryotic kingdoms), namely: protists, fungi, animals, and plants. Note that the approximate number of known species is presented for each subdivision. Source: Refs. [29, 74, 252, 253].

and it is located in the internal environment of the cell (cytoplasm). Cells of eukaryotic organisms contain a double-stranded linear (usually) DNA folded inside a membrane-bound organelle, named the nucleus (from Latin – nucleus, "kernel" or "seed"; pl. nuclei). The nuclear membrane is a controlled barrier that separates the DNA molecules from the cytoplasm (Figure 1.2). Naturally, images based on electron microscopy can best show the classic structure and the inner "frozen" dynamics of eukaryotic cells (Figure 1.2). In double-stranded DNA, a cytosine molecule from one strand and a guanine molecule from the other strand, form three hydrogen bonds while adenine and thymine form two hydrogen bonds. The successive alternation of these simple hydrogen bonds along the double-stranded DNA molecule dictates the energy required to separate the two strands and establishes the local stability of the duplex. In both eukaryotes and prokaryotes, the order of the four types of nucleotides defines the information structure throughout a DNA molecule. These structures include the well-known "genes" (Greek – geneá, "generation"). Genes are regions of different lengths, found along the DNA molecule. Broadly, gene regions are in turn accompanied by regulatory structures, such as gene promoters and enhancers. Genes are involved in transcription, namely in the synthesis of RNA transcripts. Note that RNA molecules are also polymers consisting of four types of organic molecules, namely: adenine (A), uracil (U), cytosine (C), guanine (G). The RNA transcript is a single-stranded nucleotide sequence that is complementary to the DNA strand harboring the gene. In turn, the information on the RNA transcript dictates whether the transcript becomes a functional molecule within the cell or whether it becomes a template for protein synthesis.

1.4 Chromatin Structure

In multicellular organisms, every cell type usually contains the same DNA information; however, it exhibits a different phenotype. What determines this behavior ? Double-stranded DNA molecules of eukaryotic organisms are folded and distributed into chromosomes, which take the form of chromatin (DNA, histone proteins, and non-histone proteins). The basic organization of chromatin consists of filaments made of repetitive units called nucleosomes. Each nucleosome consists of eight histone proteins (i.e. type H2A, H2B, H3, and H4) that wrap ~146 base pairs (bp) of double-stranded DNA (Figure 1.3a). Nucleosomes are connected to each other by 10–80 bp of DNA associated with linker histone H1 that wraps another 20 bp [30]. This basic form of chromatin self-assembles into higher orders of organization. Inside the cell nucleus, these higher orders of chromatin organization include the chromatin fibers, the fractal globules, and reach a final level of organization, namely the chromosomal territories [31, 32].

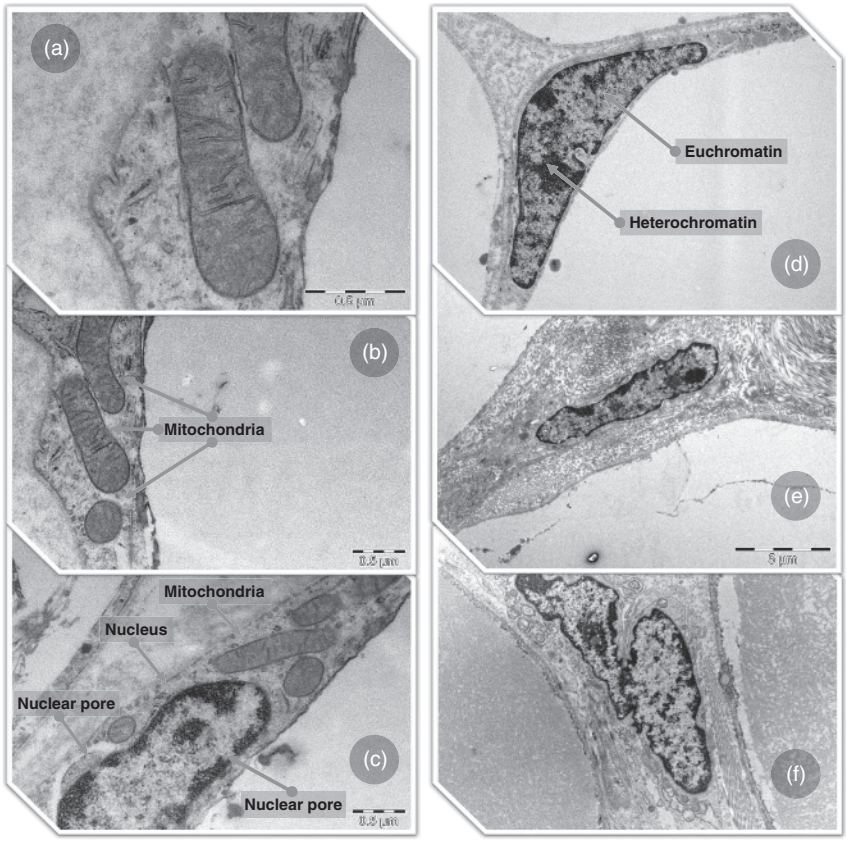

Figure 1.2 Ultrastructural images of adipocyte cells from Bos taurus (Cattles). Adipocytes especially show how tolerant and adaptable cellular organelles are to various constant mechanical stresses. (a, b) Shows mitochondria in adipocytes. The right side of homogeneous light gray content represents the lipid droplet. (c) Shows a few mitochondria in proximity to the cell nucleus. (d, e) Shows the shape of the cell nucleus in different mechanical constraints induced by the size of the lipid droplets. Again, the homogeneous light gray content represents the lipid droplets from the surrounding cells. (f) Shows two adipocytes with adjacent nuclei. Within each nucleus (c–f), the genetic material can be observed in different states of activity. Inside each nucleus, the dark gray (to almost black) areas represent heterochromatin and the normal gray areas represent euchromatin. In short, euchromatin contains a specific and dynamic set of active genes that is expressed only in adipocytes, while areas of heterochromatin contain the remaining unexpressed genes. At the edge of the nuclear membrane, nuclear pores can be observed. Interruptions with a light gray hue can be seen along the perimeter of the nuclear membrane. Those are the nuclear pores. Note that each image shows only a small fraction of the actual size of an adipocyte. Source: Courtesy Dr. Elvira Gagniuc, Department of Pathology, Faculty of Veterinary Medicine, University of Agronomic Sciences and Veterinary Medicine, Bucharest.

Chromatin is a highly dynamic three-dimensional structure, which self-arranges differently from one cell type to another, thus, establishing and maintaining the cell identity [31, 33, 34]. But what determines these chromatin self-arrangements? The predispositions for self-arrangements are determined in advance by chromatin remodelers [35]. Among the chromatin remodelers, two main families of enzymes are heavily involved in the global chromatin organization during the cell cycle, namely acetyltransferases and deacetylases. These enzymes make changes to the histone tails of the nucleosomes (histone tails are amino acid chains that extend from the nucleosomal core – Figure 1.3a). Histone tail acetylation on the nucleosomes leads to a relaxed form of chromatin, which subsequently allows transcription factors (TF) to gain access to their target genes (euchromatin). Histone deacetylation leads to a higher-order folding of nucleosomal arrays, which in turn form a dense chromatin structure that is inaccessible to the transcription machinery (heterochromatin) [36]. Thus, patterns of histone acetylation and deacetylation along the chromatin filaments dictate the initial chromatin folding and unfolding inside the cell nucleus and consequently the activity of a specific subset of genes [37]. These acetylation–deacetylation mechanisms are combined (among others) with DNA methylation mechanisms, which lead to a gradual and stable inactivation of certain genes over several cell generations (a topic discussed in other chapters). Nevertheless, the global distribution of chromatin is established immediately after the cell division and exposes a subset of genes specific to the cell type. The DNA regions that contain the genes that are part of the cell type subset are positioned more toward the center of the cell nucleus in the relaxed volume of the chromatin (called euchromatin) [31]. In contrast, genes outside the cell-type subset are distributed in the condensed volumes of chromatin (called heterochromatin) that are usually positioned toward the inner part of the nuclear membrane (Figure 1.2c–f).

Nevertheless, chromatin also undergoes partial changes while in G0 phase (the resting stage of the cell cycle – e.g. many neuronal cells are always in this state). These partial and continuous changes of the chromatin structure are meant to silence or activate certain genes from the main subset of genes. Such facultative heterochromatin areas are present on the outer surface of the heterochromatin landscape (the euchromatin – heterochromatin borders), and their condensation state depends on successive interactions between different gene products of the subset [31]. Note that DNA molecules are present not only in the area of the cell nucleus but also in other organelles (e.g. mitochondria and chloroplasts). Much like in prokaryotes, the circular double-stranded DNA molecules found in these eukaryotic organelles show their own type of organization called a nucleoid (meaning nucleus-like; it is an infrequent DNA–protein assembly) [38, 39].

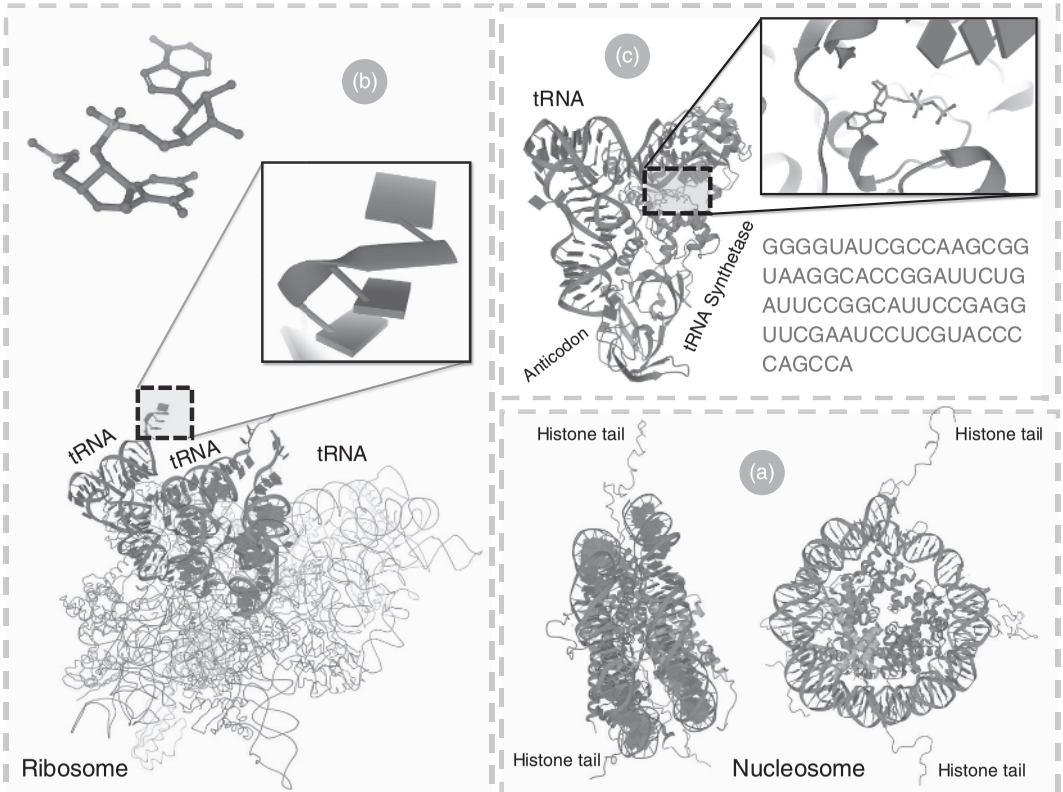

Figure 1.3 Molecular representations. (a) Shows the structure of the nucleosome core particle [52, 53]. (b) Shows the path of mRNA through the ribosome by pointing out the collinearity between the tRNA anticodons [53, 54]. The window highlights the binding region between an amino acid and a tRNA. (c) Shows the *Escherichia coli* glutaminyl transfer RNA synthetase complexed with transfer RNA(Gln) and ATP [53, 55]. The tRNA sequence is presented next to this ribonucleoprotein particle. The last letters in the sequence correspond in reverse order to the region in the tRNA highlighted by the dotted line window (i.e. "ACCG …"). The position of the tRNA anticodon is also highlighted here. Source: Refs. [52, 53, 55].

1.5 Molecular Mechanisms

Eukaryotes and prokaryotes prefer different strategies for synthesizing multiple proteins from a single DNA region (a transcription unit). In prokaryotes, several protein-coding areas (genes) are arranged linearly in a region called an operon, which is usually regulated by a single promoter. An operon is a cluster of coregulated genes with related functions [40]. Thus, operon expression leads to a number of proteins equal to the number of coding areas (genes) in the operon. All the genes in the operon are transcribed into a continuous RNA molecule, which is almost simultaneously translated into proteins. However, functional gene clustering (operon-like) has been reported in eukaryotes (i.e. fungi, plants, and animals) [40]. Eukaryotes, on the other hand, primarily use a single coding area interrupted by noncoding areas (introns). Different combinations between smaller fragments (exons) of the coding area lead to several types of RNAs and consequently to several types of proteins. The protein versions that originate from to a single gene are called "protein isoforms." Note that protein isoforms are not necessarily functionally related [41].

1.5.1 Precursor Messenger RNA

Promoter and enhancer regions regulate the transcription of nearby genes. The initiation of transcription is conditioned by various regulatory proteins that bind to the regulatory regions of DNA (the promoter and enhancer regions). In eukaryotes, the regulatory proteins facilitate changes in the local chromatin structure to allow proper recruitment and binding of RNA polymerase to one of the DNA strands. Thus, the local chromatin structure either promotes or inhibits RNA polymerase and TF binding. Transcription begins once the RNA polymerase enzyme binds to the promoter region of the gene. Regulatory proteins in conjunction with different combinations of TF dictate the frequency of synthesis for pre-messenger RNA (mRNA) molecules (how many copies per unit of time). For instance, different combinations of TFs lead to different three-dimensional macromolecular conformations (the transcription mediator complex) [42]. These temporary macromolecular constructions (made of TFs and other proteins) and their interaction with chromatin, allow the access of RNA polymerase to the DNA sequence to a greater or lesser extent. The difficulty of recruitment imposes a probability distribution for binding. In turn, this binding probability of RNA polymerase sets the frequency of synthesis for pre-mRNAs. As a rule of thumb, a more open chromatin structure is associated with active gene transcription events, while a more compact chromatin structure indicates transcriptional inactivity (no expression).

1.5.2 Precursor Messenger RNA to Messenger RNA

The DNA region of the gene can be subdivided into other types of regions. Especially in eukaryotes, many genes are organized into coding (exons) and noncoding regions (spliceosomal introns). Both exons and introns are transcribed into a continuous pre-mRNA fragment. While the pre-mRNA is being transcribed/synthesized, the intronic regions are removed by spliceosomes (a ribonucleoprotein complex) and the ligation of the exon regions forms the mRNA. The process of intron removal is called splicing. Exon ligation in the same order, in which these regions appear in a gene, is called constitutive splicing. Thus, constitutive splicing allows for a "one gene–one protein" model (or, one pre-mRNA – one mRNA). When exon ligation does not follow the order observed in the gene (i.e. certain exons are skipped), several mRNA variants are produced from a single pre-mRNA variant. If the gene encodes for proteins, then each mRNA variant will generate a different type of protein (protein isoforms). This process is known as alternative splicing, and is responsible for the exaggerated abundance of protein types in the eukaryotic proteome.

1.5.3 Classes of Introns

Introns are regions that interrupt the coding region of functional RNA or protein-coding genes. There are four known classes of introns: Group I introns, Group II Introns, nuclear pre-mRNA introns (Spliceosomal introns), and transfer RNA (tRNA) introns. Group I introns are self-splicing introns and are found in some ribosomal RNA (rRNA) genes [43]. Group II introns are mobile ribozymes that self-splice from precursor RNAs (pre-RNAs) and are found in bacterial genomes and organellar genomes, suggesting that catalytic RNAs, as informational structures, predate the origin of eukaryotes and perhaps the origin of cellular life [44, 45]. Nuclear pre-mRNA introns are found in protein-coding genes and require a ribonucleoprotein complex (spliceosomes) for splicing. The tRNA introns are found in various tRNA genes in all the three kingdoms of life, and require certain enzymes for splicing [46].

1.5.4 Messenger RNA

The stochastic behavior of brownian motion (random walk) and the concentration gradients, scatter the mRNA molecules from the origin of synthesis to other locations with a lower concentration. Thus, the mRNA scattering (as is the case with many other molecules of different sizes and shapes) is done naturally on the least frictional paths. In the case of eukaryotes, the origin of pre-mRNA synthesis and processing is the inner space of the cell nucleus (high mRNA concentration) and the mRNA molecules diffuse through the nuclear pores into the relaxed

environment of the cytoplasm (low mRNA concentration). In the case of prokaryotes, the mRNAs diffuse from the origin of synthesis (which is close to the DNA molecule floating directly in the cytoplasm) into the rest of the cytoplasm. The information from some mRNAs allows for protein synthesis, whereas the information from other RNAs provides direct biological functionality. Moreover, some mRNA molecules become functional by themselves due to a self-complementary between different regions of the same molecule. Other RNA molecules become functional after they are processed by different proteins (or vice versa).

1.5.5 mRNA to Proteins

In both eukaryotes and prokaryotes, mRNA molecules, which contain the information structure for protein synthesis, are stochastically encountered by two ribosomal subunits that initiate the translation step. Once bound to an mRNA transcript, the two subunits form the ribosome. The ribosome is a ribonucleoprotein (made of RNA and proteins) organelle that facilitates the formation of chemical bonds between amino acids in the order specified by the information encoded in the mRNA molecule. Life evolved a molecular scheme for translation, known as the "genetic code" [47]. In this scheme, groups of three nucleotides are associated with different amino acids used for polypeptide synthesis. Each set of consecutive and nonoverlapping nucleotide triplets on the mRNA transcript is known as a codon. Polypeptide synthesis begins from a start codon, which initiates the position of the reading frame. Usually, the start codon is represented by the "AUG" triplet (representation with the highest frequency across all life). However, other triplet combinations (non-AUG start codons) can take the role of a start codon (with a lower frequency) [48]. Post initialization, the mRNA transcript slides in between the two ribosomal subunits by one codon at a time following the reading frame set by the start codon [49, 50]. Different versions of tRNAs present in various concentrations in the cytoplasm are each linked to an amino acid. The type of amino acid connected to a tRNA is associated with an anticodon, a special nucleotide triplet region from the tRNA destined for a temporary bind to an mRNA transcript. Thus, tRNAs are the temporary links between the mRNA transcript and the nascent amino acid chain. An assembled ribosome contains three "openings" (A, P, and E sites) for tRNA–mRNA interactions (Figure 1.3.b). The smaller subunit of the ribosome allows for a complementary between three nucleotides (the codon) on the mRNA transcript and three nucleotides (anticodon) of a tRNA molecule (Figure 1.3.b). Once the mRNA–tRNA binding has been facilitated by the smaller subunit, the amino acid transfer from a tRNA to the nascent amino acid chain is facilitated by the larger subunit of the ribosome [51]. The tRNA molecules with appropriate anticodons come into contact through complementary with the mRNA transcript.

The amino acid chain is passed from the previous tRNA to the amino acid of the next incoming tRNA, increasing the growing peptide by one amino acid on each switch. Thus, the amino acid chain remains attached to the most recently bound tRNA and is not released until a termination codon appears in the mRNA transcript (UAA, UAG, UGA) [56]. Since it is an evolved/evolving scheme, small variations of the genetic code exist above different kingdoms of life, and these variations are central to the ultimate goals of bioinformatics (i.e. how life works).

1.5.6 Transfer RNA

On the other side of the translation, an ancient group of enzymes set the rules of the genetic code [57]. The aminoacyl–tRNA synthetase (tRNA-ligase) represents a group of enzymes. The function of these enzymes is to attach an appropriate amino acid to a corresponding tRNA (Figure 1.3.c). Many of these enzymes recognize their tRNA molecules using the anticodon [58]. Consequently, there is one tRNA-ligase for each tRNA–amino acid pair. For instance, in humans there are twenty different types of aminoacyl–tRNA synthetases, one for each amino acid of the genetic code [59]. Some organisms lack the genes needed for all twenty aminoacyl–tRNA synthetases. However, such organisms use all twenty amino acids for protein synthesis. In such cases, a tradeoff is made in the complexity of a tRNA-ligase, such that one enzyme associates more than one pair [60, 61]. Thus, the tRNA matching with an amino acid is based on additional properties exhibited by the tRNA, such as the geometry (shape) of the molecule, specific nucleotide positions along the tRNA chain, and so on [62].

1.5.7 Small RNA

RNAs have multiple and versatile roles across all biological systems and one of the roles is mRNA silencing and post-transcriptional regulation of gene expression. Small RNAs are short (~18–30 nucleotides), noncoding RNA molecules that can regulate gene expression in both the cytoplasm and the nucleus. A few classes of small RNAs have been defined, such as microRNAs (miRNAs), small interfering RNAs (siRNAs), and Piwi-interacting RNAs (piRNAs) [63]. For instance, miRNAs are small noncoding RNA molecules (~21–25 nucleotides in length) that play an important regulatory role in animals and plants by targeting specific mRNAs for degradation or translation repression [64, 65]. It appears that an imperfect complementary between miRNAs and different mRNA targets has the potential to regulate several genes simultaneously. Moreover, miRNAs cross the boundary of a single cell. To add to the complexity of these processes, some miRNAs are secreted into exosomes or microvesicles and may have the ability to move through circulation to other distant cells or tissues [66–68]. Without question, the fine-grained regulation that underlies the complexity of eukaryotes is found in these short RNA molecules.

1.5.8 The Transcriptome

The set of all RNA molecules produced by a given organism is known as the "transcriptome." This includes, of course, the mRNA transcripts but also the RNA molecules mentioned above (i.e. mRNAs, tRNAs, rRNAs, siRNAs, miRNAs, piRNAs, and so on) as well as other uncharacterized noncoding RNA molecules. When expressed, genes produce mRNAs in different quantities, which are then detectable [69]. Currently, two main techniques are representative for capturing gene expression, namely: RNA-Seq and microarrays [70, 71]. RNA-Seq (RNA sequencing) allows for full sequencing of all RNA molecules present in a sample, whereas microarrays target known transcripts of different genes through hybridization (complementary) [72]. Thus, RNA-Seq experiments can estimate the subset of genes expressed in a cell type or in different tissues (several cell types) at any one time by an alignment of the sequenced RNAs to the reference genome (the DNA of the organism) [73]. However, the transcriptome can be seen as an ideal set, because the complete set of possible RNAs cannot be fully detected. Reasoning dictates that each state of a cell shows a specific subset of RNAs from the transcriptome. Of the total number of states that a cell can exhibit, only a few states can be induced and captured by RNA-Seq. Thus, a small subset of RNAs from the transcriptome may remain undetectable. At the tissue level, there are a number of cell types, each with a specific set of active genes. Often, the analysis of the pattern of gene expression is performed at the tissue level, i.e. on several cell types at the same time. From a global perspective, this leads to a union between the sets of genes expressed in each of the cell types that make up the tissue. Furthermore, genes that are expressed in several cell types (such as housekeeping genes) may show the highest amounts of mRNA, while genes that are only expressed in certain cell types can show lower amounts of mRNA.

1.5.9 Gene Networks and Information Processing

The mRNA and/or the protein products encoded by one gene often regulate the expression of other genes. In multicellular eukaryotes, the set of genes that are expressed in a specific cell type forms an "open" gene network. Each gene network is a self-orchestrated feedback loop constantly adapting to different inputs from the environment. The dynamics of a gene network may be deduced in practice from the gene expression levels. The RNA-Seq technique shows the set of genes, and their expression levels (amount of mRNA) at the time of cell/tissue sampling. Repeated sampling at different time intervals can complete a puzzle related to the functional relationship between the genes of the set. Direct or indirect activation of a gene promoter by the product of other genes (mRNA or proteins) is done with a relative delay and largely depends on the frequency by which the gene product is synthesized. The frequency of synthesis impacts the time of accumulation of the gene product (mRNA or proteins) in the cell as well as its stochastic diffusion

toward other promoters and macromolecules with which it can interact. Note that the environment can be represented by a number of factors: the current set of molecules inside the cell, the signal molecules synthesized by other cells (other gene networks) or the amount of nutrients, pressure, temperature, and so on.

1.5.10 Eukaryotic vs. Prokaryotic Regulation

In prokaryotes, gene expression is primarily regulated at the level of transcription. Moreover, transcription and translation occur almost simultaneously in the cell cytoplasm. Eukaryotic regulation of gene expression is dynamically orchestrated at several levels, such as epigenetics (chromatin and TF), transcription, post-transcription, translation, and post-translation (further processing of the amino acid polypeptides from a primary structure to more complex, secondary, tertiary structure, and so on). Eukaryotic gene expression occurs with a delay when compared to prokaryotes, as transcription takes place within the nucleus and translation occurs outside the nucleus within the cytoplasm.

1.5.11 What Is Life?

The information in DNA molecules supports a continuous biochemical feedback inside the cell, which self-regulates according to external and internal stimuli (i.e. nutrients, signal molecules, pressure, electromagnetic radiation, and so on). Through energy consumption, this continuous process is maintained in a permanent imbalance above the "inanimate" background. From our reference system, this dynamic self-regulating biochemical feedback is considered life.

1.6 Known Species

How many species really? Unfortunately, life is not as diverse as previously believed [74, 75]. For more than 250 years, our species has cataloged all the other land and water species at our disposal, and continue to do so. Based on this census, the tree of life contains a total of 1.43 million known species, from which almost 1.42 million are eukaryotes and 12k species are prokaryotes (Table 1.1 and Figure 1.1) [74]. Most of the time, real census ruins the "feng shui" of predictions. In total, even the most enthusiastic predictions forecast between 8 million and 11 million species in existence [74, 76, 77].

However, under the heavy umbrella of uncertainty, predictions and census rarely match when it comes to the total number of species on earth. Among the cataloged species for the tree of life, animal species constitute 78% and plant species represent 15%. Out of a total of 1.4 million species, about 1.2 million species live on land and 19k live in the aquatic environment (Table 1.1).

Table 1.1 The total number of known species.

Kingdoms	Land	Water	Total
Eukaryotes			
Animals	953k	171k	1125k
Fungi	43k	1k	44k
Plants	216k	9k	224k
Protists	21k	13k	34k
Prokaryotes			
Bacteria	10k	1k	11k
Archaea	1k	0k	1k
Total	1244k	194k	1439k

The table shows the total number of known species on earth. The data are divided into two major categories on rows, namely prokaryotes and eukaryotes; and two major categories on columns two and three, namely the environment. Note that 1k means 1000 species.
Source: Refs. [74, 283].

1.7 Approaches for Compartmentalization

Compartmentalization is not a condition for life; however, a closed environment for biochemical containment might be. Most likely, unicellular species that evolved in water had fewer survival issues related to mechanics and gravity. To obtain a specialization, some of these species evolved a cooperation between internal biochemical processes rather than between individual cells. Thus, their volume and shape increased to the point where a single cell began to resemble a multicellular organism. Such unicellular organisms contain multiple nuclei, and for historical reasons are called coenocytes (Greek – coeno, "common"; cyte, "box"). Nevertheless, the virtual "cells" of these unicellular organisms are not defined by a physical barrier. This virtualization may be achieved only through controlled biochemical interactions between concentration gradients of different types of molecules (i.e. from gradients of simple nucleic acids, amino acids, fatty acids and sugars, up to RNA and proteins). Spatially spaced point sources of such chemical and biochemical gradients can form a well-organized virtual structure in these unicellular organisms. Moreover, the biochemical versatility can continue up to the point of inclusion of smaller unicellular species to form a biochemical symbiosis.

1.7.1 Two Main Approaches for Organism Formation

Life shows two main approaches for organism formation. The first approach forced a cooperation between biochemical processes (virtual cells in one physical boundary). The second approach forced a cooperation for gradual specialization among individual cells of the same species (or even between species). Interestingly, the second approach shows a lower entropy than the first. However, cooperation between individual cells did not rule out further specialization between biochemical processes inside individual cells.

1.7.2 Size and Metabolism

Competition and gravity preclude the emergence of unicellular organisms over a certain size. Moreover, gradient-based biochemical signaling and interactions would be inefficient on long distances inside large unicellular organisms. Multicellular organisms seem to have found a balance between the speed of response and the size of the cells. Small cells have a larger surface area relative to their volume. Each unit of volume can exchange gases and nutrients at a higher rate compared to larger cells. Note that the principle is equivalent to smaller salt granules that dissolve faster in water than large ones. Cooperation for development of cell specialization in the direction of a circulatory system formation ensured an optimal exchange with the outside environment and a fast response for the entire organism. In the case of very large unicellular organisms, the response time for any stimulus may be dictated by distances inside the cell and the metabolic rate. For instance, a biochemical interaction between two points in the cytoplasm of such an organism would require time and high amounts of messenger molecules to diffuse in a large volume until the target is stochastically encountered. In other words, "time contracts" for giant unicellular organisms. It is likely that giant single-celled organisms have existed in the distant past. However, competition with smaller unicellular organisms with higher response times may have eliminated them from the evolutionary chain.

1.8 Sizes in Eukaryotes

Physical sizes of different organisms provide an intuitive assessment over their complexity (Table 1.2.). The comparison between the extremes of eukaryotes can be particularly important for a good view of nonlinearity that exists between entropy and size, of which, we will discuss about later. This subchapter briefly discusses the physical sizes of different unicellular and multicellular eukaryotic organisms.

Table 1.2 Extreme sizes of unicellular organisms.

Unicellular organisms	Eukaryotes (μm)	Prokaryotes (μm)
Min	0.8	0.15
Max	300 000	1400

The table shows the minimum and maximum physical dimensions of unicellular organisms in both eukaryotes and prokaryotes. The values represent averages of the measurements published in the scientific literature and are presented in micrometers.

1.8.1 Sizes in Unicellular Eukaryotes

Eukaryotic unicellular organisms provide important clues to the schism that led to multicellularity. Also, single-celled eukaryotic organisms are particularly important biological models that can lead to a deeper understanding of the fundamental mechanisms behind the evolutionary process.

Marine life shows both the maximum and minimum sizes for unicellular organisms. For instance, a member of the green algae, *Caulerpa taxifolia*, is a unicellular organism of 30 centimeters in length, or more [78]. The *Syringammina fragilissima* is another example of a unicellular organism, which reaches ∼20–25 cm in diameter or *Ventricaria ventricosa,* which is a cell of 2–4 cm in diameter [79, 80]. On the other hand, the smallest unicellular eukaryote appears to be *Ostreococcus tauri*, a marine green alga with a diameter of about 0.8 μm [81, 82].

1.8.2 Sizes in Multicellular Eukaryotes

Through cooperation, eukaryotic multicellular organisms have been able to evolve large dimensions. In water, buoyancy counterbalances gravity and it allowed for evolution of the largest organisms on the planet. For instance, *Balaenoptera musculus* (the blue whale) is a marine mammal of 27–30 m and around 170–200 tones [83]. It may be the largest contemporary organism on the planet. On land, *Loxodonta africana* (the African savanna elephant) is the largest living land animal [84]. Among birds, *Struthio camelus* (the common ostrich) can reach 2.8 m in height and weigh over 150 kg [85].

1.9 Sizes in Prokaryotes

On the other hand, prokaryotes from *Mycoplasma* species show some of the smallest possible dimensions for life (∼100 species). For instance, bacteria

Mycoplasma gallicepticum and *Mycoplasma genitalium* are likely two of the smallest self-replicating forms of life, with a diameter of ~0.0002 mm (0.2 μm or 200 nm) [86, 87]. This small size is 2 up to four times smaller than the wavelength of a photon of light from the visible spectrum (700–400 nm). However, the largest species of bacterium found among prokaryotes are *Thiomargarita namibiensis* and *Epulopiscium fishelsoni* (between 0.5 and 0.7 mm), which are comparable in size to some unicellular eukaryotes [88, 89].

1.10 Virus Sizes

Small viruses are predominant and may be round, or rod shaped, or a combination of the two in the case of prokaryotes. A capsid is the protein shell of a virus, which can remind us of the idea of Platonic solids. For some small viruses, usually the self-assembly process is dictated by their nucleic acids sequence (motif seeds) [90]. This protein shell seals their genetic material from the environment. However, many capsid proteins can self-assemble with no additional help. Viral proteins have structural properties that allow regular and repetitive interactions among them. Identical protein subunits are distributed with helical symmetry for rod-shaped viruses and polyhedron symmetry for round viruses. Because they have small genomes, viral genes must repeat protein subunits. Each subunit has identical bonding contacts with the neighbors. Repeated interaction with chemically complementary surfaces at the subunit interfaces, naturally leads to a symmetric arrangement (3D patterns). The bonding contacts are usually noncovalent. This ensures error-free self-assembly and reversibility. Thus, if the gene responsible for the viral protein is inserted into another cell for expression, that cell will produce viral proteins that will self-assemble into shells – fake viruses with no genome inside. Depending on the species, the 3D shape and the repetitive interactions of viral proteins allow for different types of bonding patterns, which in turn lead to different configurations and capsid sizes (Table 1.3.).

Thus, some viruses can be comparable in size with certain life forms (Table 1.4.). For a degree of comparison, *M. gallicepticum* is 3 up to 10 times smaller than the

Table 1.3 Extreme sizes in viruses.

Viruses	Eukaryotic viruses (μm)	Prokaryotic viruses (μm)
Min	0.017	0.03
Max	1.5	0.2

The table shows the minimum and maximum physical dimensions of viruses found in both eukaryotes and prokaryotes. The values represent averages of the measurements published in the scientific literature and are presented in micrometers.

Table 1.4 Single-celled organisms vs. viruses.

	Eukaryotes		Prokaryotes		Eukaryotic viruses		Prokaryotic viruses	
	Species	**Val (µm)**	**Species**	**Val (µm)**	**Species**	**Val (µm)**	**Species**	**Val (µm)**
Min	*Prasinophyte algae*	0.8	*Mycoplasma genitalium*	0.15	*Porcine circovirus*	0.017	Phages	0.03
Max	*Caulerpa taxifolia*	300 000	*Thiomargarita namibiensis*	1400	*Pithovirus sibericum*	1.5	Phages	0.2

The table shows a comparison between extreme microscopic sizes of viruses and unicellular organisms, that covers both eukaryotes and prokaryotes.

diameter of the largest giant viruses. Giant viruses that infect single-celled eukaryotes like amoebas (i.e. *Acanthamoeba castellanii*), such as *Pithovirus sibericum*, *Pandoravirus salinus,* or *Pandoravirus dulcis*, are about 1–1.5 μm (1000–1500 nm) in length [91, 92]. Other more well-known giant viruses are *Megavirus chilensis* (400 nm) or *Acanthamoeba polyphaga Mimivirus* (390 nm), each with considerable dimensions over the size of certain prokaryotes [91]. In terms of physical size and genome complexity, giant viruses closed a significant gap between the realms of viruses and the prokaryotic/eukaryotic unicellular organisms [91].

On the other scale, *Porcine circovirus* is the smallest virus (17 nm) and is found in multicellular eukaryotes [93, 94]. Almost all isolated viruses from prokaryotes show ranges between 30 and 60 nm. Giant prokaryotic viruses with capsids diameters ranging from 200 to more than 700 nm have been reported [95]. Nevertheless, these comparisons between virus sizes in prokaryotes and eukaryotes can be misleading as more specialized life forms can lead to more extreme variations in size, complexity, and methods of infection.

1.10.1 Viruses vs. the Spark of Metabolism

How can *P. sibericum* be so big yet lifeless? There are several reasons for which viruses are not considered alive nor do they become alive from our perspective. More robust viral species of considerable size have a reasonable probability to incorporate parts of biochemical mechanisms from the infected cells (inside their capsid). Thus, although giants viruses may incorporate functional metabolic pathways of a cell, those functional parts will have nothing to consume since the capsid does not allow the proper exchange of molecules between the interior of the capsid and the outside environment. Those metabolic pathways that can consume component parts inside the capsid may inactivate the virus. Even assuming that there can be a possibility for a primitive metabolism, capsid proteins hinder replication of a possible "new life form." This is the likely reason why a virus of considerable size lacks the spark of metabolism. But are viruses alive? The virus environment is the cell. Without this environment, viruses become inactive until different stochastic processes lead to reactivation. For cells, the environment is represented by molecules that can be metabolized. Without these substances, cells either decay in simpler macromolecules or enter a hibernation state like viruses do. Therefore, the answer is relative and dependent on our reference system.

1.11 The Diffusion Coefficient

But why a discussion about the size of organisms? Mass diffusivity (diffusion coefficient) is a physical constant that impacts the way an organism can evolve.

The cell volume must be balanced with the cell surface; otherwise, the exchange with the external environment becomes inefficient. This exchange consists of metabolites that must exit the cell per unit time or nutrients that must enter the cell per unit time. Multicellularity allows an organism to exceed the size limits normally imposed by diffusion. On the other hand, unicellular organisms with increased size have a decreased surface-to-volume ratio and may have difficulty in absorbing and transporting sufficient nutrients throughout the cell. As a counterbalance, unicellular eukaryotic organisms have among the most varied shapes and sizes observed in nature. Both unicellular and multicellular organisms can achieve a high surface-to-volume ratio by favoring DNA mutations that lead to a convoluted surface. For instance, to increase their surface area, choanocyte organisms can take many forms, such as *C. taxifolia,* which resembles a kind of "pine leaf" or *S. fragilissima,* which has a convoluted surface.

1.12 The Origins of Eukaryotic Cells

All eukaryotic cells contain membrane-bound organelles (e.g. the nucleus, mitochondria, chloroplasts, and so on). The complexity of the eukaryotic cell is given by the presence and the interaction of organelles. The origin of organelles has always been a mystery difficult to explain. However, the endosymbiotic theory is the leading evolutionary theory for the origin of eukaryotic cells. The idea of endosymbiosis was first proposed by Konstantin Mereschkowski in 1905 [96, 97]. According to the theory of endosymbiosis, the eukaryotic cell is like a Matryoshka (Russian doll). A symbiotic relationship where one organism lives inside the other is known as endosymbiosis. The term "primary endosymbiosis" refers to the engulfment and retention of a prokaryote organism by another prokaryote or eukaryote organism. The term "secondary endosymbiosis" refers to one eukaryote organism having engulfed and retained another eukaryote organism with an organelle already obtained by primary endosymbiosis. Note that today the endosymbiotic behavior is most beautifully observed in protists (e.g. *Paramecium bursaria*).

1.12.1 Endosymbiosis Theory

The last universal common ancestor (LUCA) was likely a population of unicellular organisms that led to the emergence of two domains in prokaryotes: Bacteria and Archaea. It seems that LUCA were complex unicellular organisms and not the immediate descendants of primeval cells. Rumor "has it" that prokaryotes are the descendants of LUCA by reductive evolution [98]. Nevertheless, evidence shows that about 2 billion years ago, eukaryotic cells may have evolved from a

merger between the two prokaryotic domains. Endosymbiosis theory suggests a scenario in which an archaeal cell engulfed a bacterial cell. This kind of merger was repeated independently many times and eventually evolved to form all the membrane-bound organelles, including the mitochondria and chloroplasts. The last eukaryotic common ancestor (LECA) was likely a population of unicellular organisms that eventually (i.e. within 300 million years of LECA) led to a diversification of eukaryotes in supergroups (around 1.5–2 billion years ago) [99, 100]. Today, existing prokaryotic groups reveal the intermediate steps in the eukaryotic-cell evolution [101]. Moreover, complex archaea that bridge the gap between prokaryotes and eukaryotes have been found [102].

1.12.2 DNA and Organelles

Prokaryotic organisms of the distant past are perhaps the ancestors of almost all membrane-bound organelles that are found today inside the eukaryotic cells. Among the membrane-bound organelles, some contain their own genome and others lost their genome throughout evolution [103]. The adaptation to the intracellular environment led to the loss of many of the original genes accumulated for environmental survival. Other important genes from the organelles have been transferred to the nucleus over the evolutionary timeline. Organellar DNA transfer to the nucleus is a known process by which, during evolution, some critical genes of the organelles are moved for preservation and synchronization of cell division [104, 105]. But why preservation ? The DNA mutation rate is lower in the nucleus. In some important organelles, high concentrations of reactive oxygen species (ROS) can lead to oxidative stress and DNA damage [106, 107]. Thus, the relocation of a gene from the organelle to the nucleus enables a more secure conservation over time. Moreover, the transfer process also leads to an obligate codependency. The relocated genes control the division of the organelles (synchronization) and encode products that interact with organelle-encoded proteins. In turn, the genes still present in the organellar genome encode proteins that interact with nuclear proteins [108]. Ultimately, the organelle interacts with its own evolved genes physically present in the nucleus of the cell. Nevertheless, some of these DNA containing organelles still are organisms in their own right; genetically equipped for the environment imposed by the cytoplasm of the host cell. Intracellular signaling pathways that coordinate gene expression between organellar and nuclear genomes are highly complex; toward almost complicated [109]. Moreover, additional signaling pathways exist between different organellar genomes [109]. These complications may be one of the reasons for the reductive evolution of the organellar genomes. However, many organelles retained a large part from their ancestral genome. Thus, different equilibrium states between organellar and nuclear genomes must exist. Moreover,

contrary to certain sedimented expectations, prokaryotic organisms also may contain organelles with special arrangements; for instance organelles such as magnetosomes, chlorosomes, pirellulosomes, anammoxosomes, carboxysomes, and so on [110–112].

1.12.3 Membrane-bound Organelles with DNA

Mitochondria arose approximately 2.3–1.8 billion years ago from a unicellular organism related to modern α-proteobacteria. Bacterial species *Rickettsia prowazekii* of the genus *Rickettsia* is probably the closest phylogenetic relative to the mitochondria [113, 114]. Also, ATP production in *Rickettsia* is the same as that in mitochondria. On the other hand, plastids (i.e. chloroplasts and chloroplast-like organelles) arose 1.6–1.5 billion years ago from the ancestors of cyanobacteria [115]. That is, a mitochondriate eukaryote became host to a cyanobacterium-like prokaryote. Organelles with their own genomes, such as plastids and mitochondria, are found in most eukaryotic cells [116]. A multicellular eukaryote contains hundreds to thousands of these organelles in each cell. The number of organelles is specific to each cell type and may vary depending on the state/metabolic needs of the cell.

1.12.4 Membrane-bound Organelles Without DNA

How much of the genome can be transferred to the nucleus or can be permanently lost in evolution? The complete loss of DNA in an organelle is a possibility. Among the organelles that have lost their genome in evolution is the hydrogenosome. Hydrogenosomes are cell organelles that have a double membrane and synthesize ATP via hydrogen-producing fermentations [117]. Hydrogenosomes were once mitochondria and are a classic example of complete mitochondrial genome loss [118]. Considering the hydrogenosome example, it is reasonable to assume that all eukaryotes may contain an organelle of mitochondrial ancestry [119]. However, DNA in the hydrogenosomes of some anaerobic ciliates has been detected [120]. Ciliates are protists with hair-like organelles (i.e. cilia) used for propulsion and adherence inside liquid media. Cases of genome-containing hydrogenosomes show that these organelles are somewhere toward the end of their reductive evolution period. It can also mean that the loss of the genome is not just a one-way street, which would have important implications for elucidating the occurrence of life on Earth. A basic question arises when considering the above: Are the genome-less hydrogenosomes organisms in their own way? This is an interesting question because it shows the versatility of life and the ideas discussed in the "*Philosophical transactions*" chapter or in the "*Viruses vs. the spark of metabolism*" subchapter.

1.12.5 Control and Division of Organelles

Both genome and genome-less organelles divide. But how ? Regular bacterial fission (division) uses a dynamin-related protein (DRP) to constrict the membrane at its inner face [121]. However, DRPs are also essential for mitochondrial and peroxisomal fission [122]. Fission is required to provide a population of organelles for daughter cells during mitosis. In contrast to bacterial fission, mitochondria use dynamin and DPRs to constrict the outer membrane (a ring) from the cytosolic face [121]. For instance, the unicellular eukaryote *Trichomonas vaginalis* is a parasite that uses hydrogenosomes instead of regular mitochondria. The role of DRPs in the division of the hydrogenosome is similar to that described for peroxisomes and mitochondria [123]. Moreover, plastids use similar mechanisms and a plant-specific DPR [124, 125]. In eukaryotes, dynamin and DPRs have their genes stored in the nuclear genome. Thus, control over the division of membrane-bound organelles is held by the nuclear genome. Genes encoding DPRs were once present in the symbiotic α-proteobacterium ancestors of mitochondria or in the symbiotic cyanobacterium ancestors of the plastids. In other words, the alternative self-assembly of a complementary dynamin constriction mechanism on the outer membrane, allowed a transition of the organellar fission genes to the nuclear genome for synchronization of cell division. Moreover, this complementary mechanism allowed a reductive evolution up to the point of complete genome elimination in some organelles, such as in the case of hydrogenosomes. But how do organelle genes physically get into the nucleus to recombine with chromosomal DNA? Insertion of DNA from organelle genomes into the nucleus is DNA-mediated (RNA-mediated insertions may also occur) through a process called HGT [126]. As a side note, peroxisomes are single-membrane organelles that catalyze the breakdown of very-long-chain fatty acids through beta-oxidation [127]. Peroxisomes are a hybrid of mitochondrial and ER-derived (ER – endoplasmic reticulum) pre-peroxisomes [128].

1.12.6 The Horizontal Gene Transfer

The symbiosis between organisms is not possible without the HGT. The HGT is the weak “glue” that unites all species and it has important evolutionary implications [129]. In the future, these implications may very well undo the classification discussed above for the tree of life and how we understand the origin of life on Earth. The HGT refers to the transmission of DNA between different genomes, whereas the vertical gene transfer (VGT) is made between generations by sexual or asexual reproduction. The way in which the classification for the tree of life works is largely based on the VGT concept; thus, one can imagine the issue. HGT was first observed as a phenomenon in *Streptococcus pneumoniae* species by Frederick Griffith in 1928 [130]. The main observation made by

Frederick Griffith was that virulence (pathogenicity) in this species of bacterium is transmitted by contact or proximity. This was an important revelation for the later field of genetics. Since then, increasing evidence shows that DNA fragments of different sizes may be exchanged between the kingdoms of life, to a greater or lesser extent [129]. Not long ago, the transfer of genetic information from the members of the *Agrobacterium* genus to eukaryote cells was seen as an extraordinary and rare process [131, 132]. Today, evidence indicates clearly that transfer of genetic information between species and inside different cell compartments is a common process, which takes place over the evolutionary time. For instance, bacteria have acquired genetic material from eukaryotic hosts and vice versa [133]. Viruses contain genes derived from their eukaryotic hosts and vice versa [134]. In plants, for instance, the HGT between genomes takes place through intracellular transfer of DNA among the nuclear, mitochondrial, and plastid genomes. The transfer of mitochondrial genes to the nucleus is known to be an ongoing evolutionary process. However, evidence also shows a HGT of mitochondrial DNA to the plastid genome [135]. Moreover, expression of a transferred nuclear gene in a mitochondrial genome was also observed [136]. For example, the orf164 gene in the mitochondrial genome of Arabidopsis is derived from the nuclear ARF17 gene that codes for an auxin-responsive protein [136]. Thus, the transfer of DNA segments from any location to any other location seems to be a rule across all life. However, HGT is most frequent between closely related species with similar genome features and less frequent otherwise [137]. In other words, HGT is a process that occurs at different frequencies between prokaryotes, between eukaryotes, between prokaryotes and eukaryotes and vice versa [138]. Perhaps, the importance of HGT goes as far as the emergence of new species (speciation) [139, 140].

1.12.7 On the Mechanisms of Horizontal Gene Transfer

Understanding of the mechanisms and vectors underpinning HGT across the kingdoms of life is still limited. Mobile genetic elements (MGEs) represent the main known vectors for HGT [137]. Well-known HGT events often include, but are not limited to, transposable elements (TE), plasmids or bacteriophage elements [141]. The behavior of a MGE has a certain degree of stochasticity and may incorporate a complete gene(s) or may include only sections of a gene, or with a high probability none of the two. Sections of genes transferred by MGEs decay in time and are recognized in bioinformatic analyzes as pseudogenes (nonfunctional genes) [126]. Among the MGEs, TE can best show the level of complexity that a DNA fragment can exhibit. The TE were first observed in *Zea mays* (corn) by Barbara McClintock in 1950 [142]. The main observation made by Barbara McClintock was that the genetic material can jump from one place to another within a genome. The

insertion of TEs into the coding pigment-genes was responsible for unstable phenotypes on the kernels of a maize ear (kernels of different colors). Note that each kernel is an embryo produced from an individual fertilization and one ear of corn contains around 800 kernels positioned in 16 rows.

1.13 Origins of Eukaryotic Multicellularity

Above the evolutionary time, cells of multicellular organisms evolved a series of states (cell types). The mechanisms that lead to the formation of such states are unknown. Biology includes several competing hypotheses on the origin of eukaryotic multicellularity; all of them based on observations made on the behavior of current species. These hypotheses suggest multiple pathways that can lead to multicellular organisms; some pathways more successful than others. Moreover, these competing hypotheses may all be valid. Note that only a few general notions are mentioned here.

1.13.1 Colonies Inside an Early Unicellular Common Ancestor

One of the hypotheses for multicellularity suggests a repeated division of the nucleus within the same unicellular organism and a subsequent formation of membranes in between the nuclei. A reminiscent coenocytic behavior can be seen in multicellular eukaryotic organisms, for instance, in the eggs (0.51 ± 0.003 mm) laid by the well-known *Drosophila melanogaster* (vinegar fly). The initial stages of the vinegar fly eggs contain multiple nuclei in a common cytoplasmic space (the entire volume of the egg) [143]. Only a few stages of development later, the cell membranes around the floating nuclei start to appear almost simultaneously to constitute the initial cells of the larva [143].

1.13.2 Colonies of Early Unicellular Common Ancestors

A second hypothesis for the origin of multicellularity proposes that unicellular organisms may aggregate to form unitary colonies that can achieve multicellularity and cell specialization over time. According to this theory, multicellularity emerged from cooperation between unicellular organisms. Examples of cooperation among organisms have been observed in nature at different scales and in various forms. One of the simplest integrated multicellular organisms is *Tetrabaena socialis* in which four identical cells constitute the individual [144]. The nuclear genome of *T. socialis* dictates the number of cells in the colony [145]. Another example is the choanoflagellate (Greek and Latin – khoánē, "funnel"; flagellate, "flagellum") *Salpingoeca rosetta,* which can exist as a unicellular

organism or it can switch to form multicellular spherical colonies called rosettes (form bridges between cells by incomplete cytokinesis), showing a primitive level of cell differentiation and specialization. Formation of multicellular colonies is induced by different signal molecules. The source of such signal molecules can originate from individuals of the same species (i.e. slime molds) or from individuals of different species (i.e. bacterium species) [146]. In the case of *S. rosetta*, the signal molecules for colony formation originates from the food source, namely the *Algoriphagus machipongonensis* bacterium (phylum *Bacteroidetes*) [147, 148]. Choanoflagellates, sponges and algae of the genus Volvox are more complex examples of first evolutionary stages that indicate the border between colonial organisms and multicellular organisms. Choanoflagellates are the closest relative of metazoans (all animals composed of cells differentiated into tissues and organs) [149, 150]. Some genes required for multicellularity in animals, such as genes for adhesion, genes for signaling, and genes for extracellular matrix formation, are also found in choanoflagellates [151]. This suggests that these genes may have evolved in a common ancestor before the transition to multicellularity in animals [152]. Sponges are one of the oldest primitive multicellular organisms in the fossil record. Choanoflagellates are small single-celled protists, partially similar in shape and function with some of the sponges cells (choanocytes) [153]. Many associations have been made in the past between choanocytes and choanoflagellates. However, the transcriptome of sponge choanocytes is the least similar to the transcriptomes of choanoflagellates and is significantly enriched in genes unique to either animals or sponges alone [154]. Slime molds are also interesting examples, which can indicate how some multicellular organisms formed. Slime molds are unrelated eukaryotic organisms that can live as single cells. In certain conditions (i.e. starvation), single cells of the same species can aggregate to form multicellular reproductive structures [155]. For instance, the multicellular aggregate (a slug-like mass of a few thousand cells called a *grex*) of amoebae *Dictyostelium discoideum* can show cellular adhesion, cellular specialization, tissue organization, and coordination that allows for mechanical movement [156, 157]. Although the behavior of *D. discoideum* is not necessarily a close example of the process that led to multicellular organisms, it can certainly serve as a clue for detailed research on the emergence of multicellularity.

1.13.3 Colonies of Inseparable Early Unicellular Common Ancestors

A third hypothesis for the origin of multicellularity suggests an early unicellular organism that underwent repeated divisions with incomplete separations between generations, which further led to a forced cooperation and specialization for a primitive tissue formation. Plant embryos and animal embryos adhere to this behavior. More to the point, the eukaryotic microorganism *Saccharomyces*

cerevisiae (the budding yeast) is an ideal example for this hypothesis [158]. Yeast colonies growing on solid media show specific structural patterns in their three-dimensional multicellular organization. These structural patterns are specific to each yeast strain [159]. Moreover, variations in the multicellular organization appear to be dependent on the environmental parameters, such as the position of surrounding cells, nutrient gradients, temperature, and so on [160].

1.13.4 Chimerism and Mosaicism

The cooperation of eukaryotic cells is best observed in the case of two phenomena, namely chimerism and mosaicism. Mosaicism is represented by two or more cell populations in different tissues originating from one fertilized egg. Namely, one cell population with the original genotype (usually representing the majority) and other cell populations with slight variations of the original genotype. One of the mechanisms that lead to mosaicism is represented by transposomes [161]. With embryonic development, the genotype of an organism can undergo various types of mutations, including transposome-induced mutations above different cell lines. These mutations can occur late in embriogenesis, leading to marginal effects at the organism level, or they can occur early in the embryonic development of an organism with more pronounced/noticeable effects [162]. Transposome-induced mutations represent a normal and nonrandom variability in multicellular organisms, leading to different phenotypic characteristics [161]. The classic example, however, is represented by the experiment performed by Barbara McClintock on corn kernels, where the transposomes inactivate the gene for the pigment protein and the phenotype is easily recognizable (please see the "horizontal gene transfer" subchapter from above). Mosaicism can also be represented by other types of mutations. For example, in humans, Down syndrome is characterized by an additional copy of chromosome 21, which is frequently attached to chromosome 14 (Trisomy 21). The extra copy of chromosome 21 slightly changes the chromosomal territories in the cell nucleus and the way heterochromatin and euchromatin are distributed. This leads to unusual variations in the expression of certain genes, especially those present on the extra chromosome 21 and those present on the neighboring chromosomal territories. Trisomy 21 occurs at the beginning of embryogenesis. However, such a mutation may appear late in embryogenesis, which results in mosaic tissues, part with normal cells and part with cells with an extra copy of chromosome 21 [163, 164]. Such a mosaic can be clinically unnoticeable unless genetic analysis is made on different tissues of the organism. Moreover, it is believed that a combination of germinal and somatic Trisomy 21 mosaicism may be reasonably common in the general population [163]. In development, cells with different genotypes compete in the tissue population. Such a competition can

lead to the possibility (especially for mosaicism that appeared late in embryogenesis) in which cells with the original genotype completely marginalize the function of mosaic genotypes or vice versa, depending on which is more fit for a specific function. On the other hand, chimerism is represented by fusion of more than one fertilized zygote, namely cells of different organisms that are orchestrated by common molecular signals to form a single body [165]. Chimerism can be observed in all multicellular species to a greater or lesser extent and may be accompanied by genetic mosaicism in any of the genotypes that form the composite organism.

1.14 Conclusions

Biological literature is probably the most sophisticated among all sciences and can be particularly overwhelming. An introduction was made to some important concepts that can provide an overview on living organisms, such as the emergence of life, classification, number of species, the origins of eukaryotic cells, the endosymbiosis theory, organelles, reductive evolution, the importance of HGT, and the main hypotheses regarding the origin of eukaryotic multicellularity. Among the biological concepts described here, some have wider implications. Examples of genome-less organelles, such as hydrogenosomes, or processes such as the HGT, question life as we understand it. Endosymbionts best explain the significance of the environment and also explain the distribution of life in a blurry, nonunitary context. In other words, endosymbiosis widens the threshold of life and shows how difficult it is to place a border between how much life resides inside or outside the cell. Moreover, the HGT appears to connect all the species on earth to a greater or lesser extent. Much evidence shows that some of these ancient processes (e.g. catalytic RNAs) are likely adding or subtracting innovative mechanisms for continuous adaptations among different species (if not all).