1

THE VIROSPHERE

1.1 Deep Microspace Field

The first image shown in this chapter (figure 1.1a), on the left side, is a deep field image from the far universe taken by the Hubble telescope. It was taken in 1995 and it covers an area of just about one 24-millionth of the whole sky, supposedly devoid of stars. And yet, the resulting picture, filled not with a few stars but many whole galaxies, takes your breath away. Almost all objects are galaxies that emerged in the early stages of the universe. The image next to it also recalls the deep universe. Is this picture an example of a cluster of stars or galaxies? Despite the similarities, the picture has been made using a special technique known as epifluorescence microphotography and deals with a vastly smaller scale: a small area of a drop of sea water. Even such a small amount of matter includes a huge number of microbial organisms. And here, too, scientists found much more than they would have ever suspected. The tiniest bright spots are viruses, followed in size by bacteria and archaea cells (medium-sized, around 0.5 μm in size) and also a few larger spots associated to protozoan organisms. These are no astronomic objects and yet an entire living universe inhabits the water drop.

Figure 1.1. Large and small universes. The left picture is an image from the Hubble Deep Field Team and NASA. The right picture involves a very small, living system, namely the small-scale ecology contained in a sea drop of water, where marine viruses, bacteria, and protists appear as small, medium, and large bright spots, respectively. Adapted from Fuhrman (2009).

As it occurred with the discovery of the real dimensions of the universe, which was limited to our Milky Way until the use of powerful telescopes in the 1920s (allowing us to see that ours was just one galaxy among many), novel sequencing techniques uncovered a vast hidden microbial diversity in the oceans, from insects and other invertebrates and from non-cultivated plants, all of which were poorly sampled before. As more and more microbial diversity was uncovered, a feeling built in the community that it may be just the tip of the iceberg (DeLong 1997). The rise of metagenomics[1] revealed a completely unexpected, almost astronomic diversity of marine viruses. These viruses were virtually out of the ecological description of ocean life until then, but we know now that—like dark matter in our universe—this hidden part of the living biosphere turned out to be essential to actually understanding how the biosphere works. The presence of viral diversity and its importance is highlighted by the observation that every new sequenced virome includes new sequences (Angly et al. 2006; Kristensen et al. 2010; Simmons 2015; Roossink et al. 2015; Brum et al. 2015). It can be said that the perceived relevance of marine viruses (and the biosphere in general) has moved from anecdotal evidence to the fact that they are the most abundant biological entities, amounting to 10 to 10^2 viruses over cells.

These numbers are indeed impressive and since long before the discovery of marine viruses we have been dealing with these entities through our evolutionary history. As we will discuss in the following chapters, viruses have played a major role not only in our evolutionary but also in our recent history. They have been shaping our genomes and physiological features in surprising ways. They can be deadly or good. They can change

[1]Metagenomics consists of the characterization of genomic materials from environmental samples. By using advanced nucleid acid-sequencing techniques it is now possible to characterize microorganisms present in the samples without the need of isolating and cultivating them.

so fast and affect so many cellular pathways that mounting evidence supports a picture of evolution as largely dependent on the driving force provided by these entities. Their importance if so great that a "virocentric" perspective of evolution cannot be avoided (Koonin and Dolja 2013). Viruses as apparently innocent as those causing flu have killed millions of humans (and many other species), becoming a threat to our survival. But without them many major evolutionary events would have never happened.

Some viruses' names have become popular in the media because of their terrifying impact through deadly pandemic events. Two of them in particular are on top of our list: the *Human immunodeficiency virus type 1* (HIV-1, figure 1.2) and the Ebola virus (EBOV, figure 1.3). They both illustrate the simplicity that can be achieved by viral agents, equipped in most cases by small genomes where a few genes coding for the essential components and copying machinery often overlap in order to maximize information compression. Both examples involve small structures, high mutation rates and a huge capacity to trigger fear. But they differ in many ways. The molecular logic of their replication, their repertoire of cellular targets to infect, their origins, and how they spread through human populations are different.

HIV-1 became known as a real threat once new cases of a previously unknown disease started to become common in the 1980s. No one would have suspected then that the new pathogen would spread around the entire planet and produce a great pandemic killing tens of millions of individuals. It became known at some point that the virus was hiding inside cells and that carriers were free of symptoms over a long time interval before a collapse of the immune system occurred with fatal consequences. During this silent period, infections would occur and the virus propagated exponentially. HIV-1 spread through both rich and poor countries. Until its biology was well understood and its

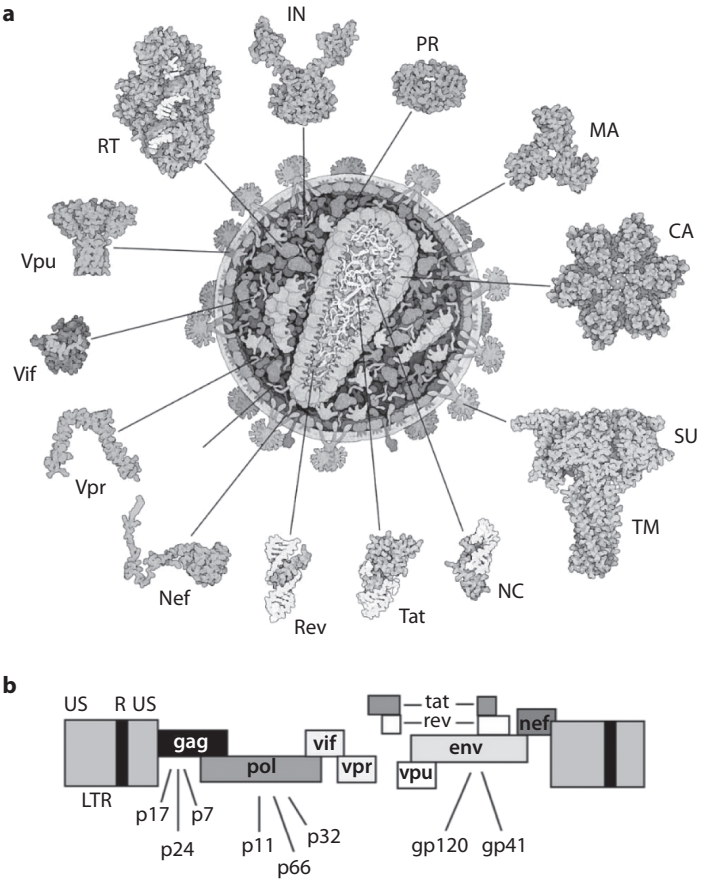

Figure 1.2. A well-known example of an RNA virus that has been responsible for one of the worst pandemics ever: HIV-1. In (a) we display the basic structure of the whole virus (central figure), and all the key molecular components are also indicated (image adapted from Goodsell (2012)). In (b) the HIV-1 small genome is schematically shown, involving just three structural genes (*gag*, *pol*, and *env*) along with regulatory elements.

Achilles' heel was discovered, the death toll grew over time. Only scientific and clinical research, including a great deal of modeling effort, eventually enabled us to properly fight back.

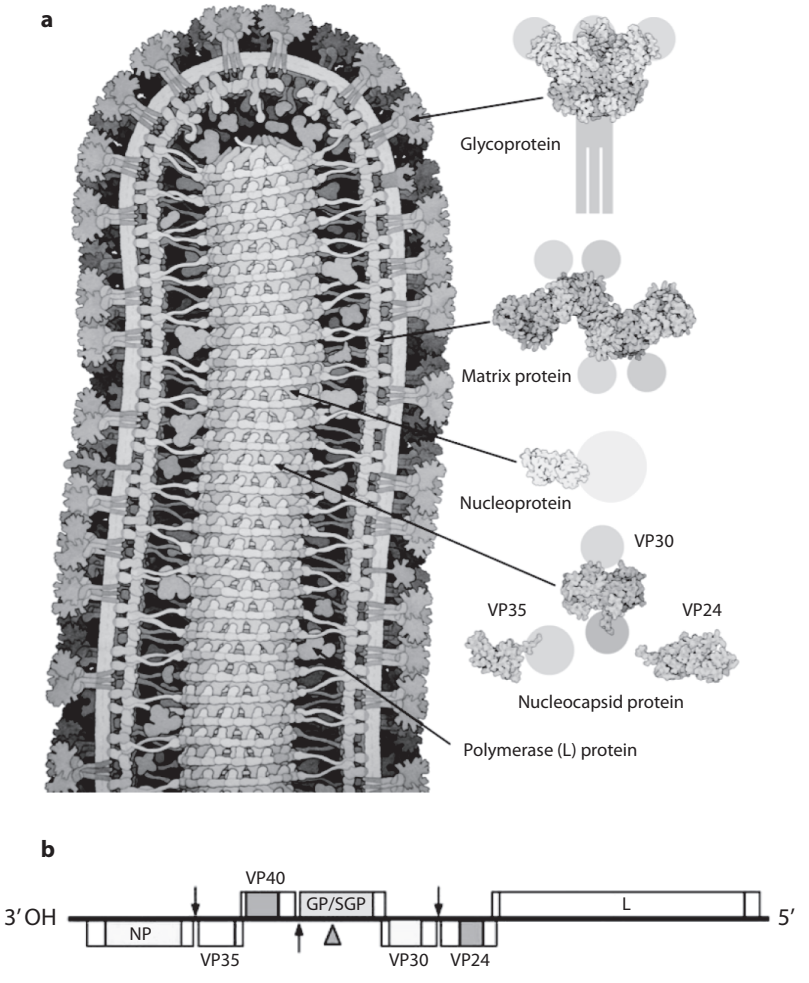

Figure 1.3. EBOV (a) an emergent virus that has likely jumped from bats to great apes and humans. It belongs to the family of filoviruses and has a characteristic filamentous shape. The key molecular components are also displayed (image adapted from Goodsell (2012)). (b) Schematic representation of the genome, encoding for seven structural and one nonstructural protein.

EBOV, on the other hand, represents a good example of another emergent pathogen notable for the bloody and deadly way in which it interacts with the human host. But in this case, the rapid damage caused to the patients prevents the virus from spreading on a global scale. However, poverty, reduced investment in healthcare, and some cultural factors have to be blamed for most EBOV outbreaks.

1.2 The Expanding Viral Universe

The impact of this hidden *virosphere* on ecosystem functioning can be summarized by means of some basic numbers (Suttle 2005; Weitz 2017). The number of viruses that might be present in the entire marine biota is 10^{30} (a 1 followed by 30 zeros) and the number of infection events taking place every second would amount to no less than 10^{23}. As a consequence of infections, viruses kill around 20% of the total microbial biomass in a single day, thus forcing a constant and large-scale population turnover. Since the microbial component of the marine biota is responsible for a major fraction of energy flows, the obvious consequence is that large-scale ecological processes are strongly constrained by the viral component of the biosphere. Marine viruses illustrate one of the most obvious results of ecological research: the realization that our planet is dominated by microbes and, very especially, by viruses.

In figure 1.4 we summarize this dominance using two main quantitative measures: total biomass and population abundance in marine communities. The biomass is clearly dominated by bacteria, with prokaryotes and viruses following closely in small fractions. However, the total number of individuals clearly differs from what is represented in the biomass picture (figure 1.4 left). Here viruses greatly outnumber other taxa, consistently with our previous picture. As pointed out by Koonin and Dolja (2013), it

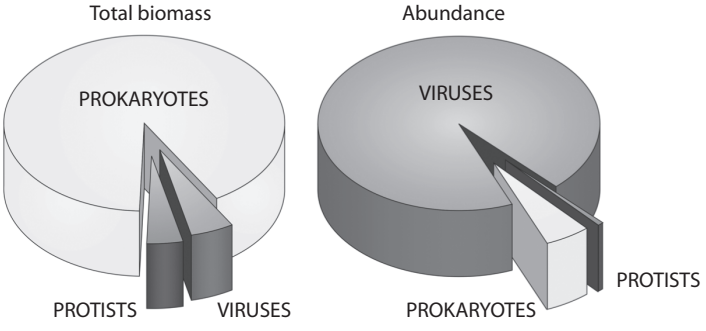

Figure 1.4. Viruses are by far the most abundant biological entities in the oceans, comprising approximately 94% of the nucleic acid-containing particles, but they only amount for 5% of the total biomass. By contrast, even though prokaryotes represent less than 10% of the nucleic acid-containing particles, they represent more than 90% of the biomass (diagram adapted from Suttle (2007)).

can be claimed that the water in the ocean is literally a virus soup with up to 10^9 viral particles per milliliter.

A very different but not less rich facet of the viral universe plays a crucial role in our own bodies. It is well known that a human needs to be seen not as an isolated entity carrying around 20,000 genes, but instead as a complex consortium of species. In particular, we are the carriers of a vast ecological web of interactions that take place among the many species of microorganisms that colonize our mouth, lungs, gut, or skin. This is known as the *microbiome*. The microbial part of ourselves carries around three million additional genes and has been coevolving with us for millions of years (Boulang and Nagler 2016; Wesemann and Nagler 2016; Taur and Pamer 2016). After the recognition of the major impact of the microbial part of our nature, the so-called *virome*, a no less interesting problem has to do with the inevitable role played by the microbe's parasites (Minot et al. 2016).

The example used above provides just a first glimpse of the enormous relevance of viruses. In this chapter, we will provide an overview of the complexity of the virosphere, addressing several key questions: What is this virosphere made of? How have viruses so successfully expanded over every single scale from bacteria to humans and even to other viruses? Is this virosphere very diverse? These questions will help to define the vast scenario that we plan to explore in this book. It is not only an extraordinary example of our biosphere's complexity; viruses themselves are a rich, and sometimes unexpected, instance of complex systems that perfectly illustrates the tempo and mode of complexity evolution and how it pervades nature.

1.3 Structural and Genetic Diversity

Viruses inhabit a domain of size ranges that spans the broad interval between molecular structures and cells. Some viruses are so small that it took a long time to detect them (figure 1.5). They were first reported in 1892 by Dmitri Ivanovsky, a Russian scientist who was studying the process of transmission of a tobacco disease. Some unknown pathogen was damaging the plant tissues and in an experiment he filtered a suspension of infected tissues through a ceramic filter, which was known to retain bacterial cells. Once filtered, the suspension was free of bacteria and yet capable of infecting, indicating that a smaller class of biological agent was responsible for the disease. In this way, the first virus was discovered: *Tobacco mosaic virus* (TMV). Other scientists independently confirmed the existence of this new class of entities, and the invention of the electron microscope was, along with the development of molecular genetics, an especially important breakthrough, since those invisible agents became visible and their internal structures and genetic components became available to study.

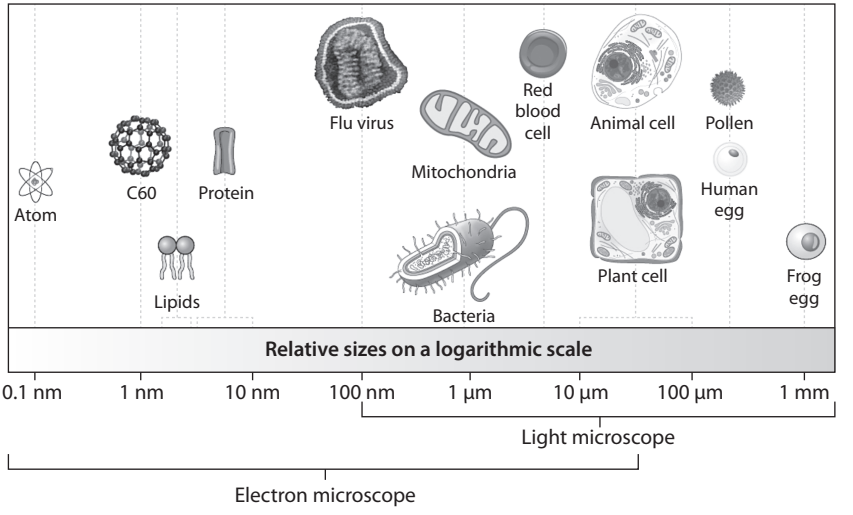

Figure 1.5. Viruses occupy an intermediate position between the macromolecules and living cells and organelles. Most of them are 10^2 times smaller than the smallest cells, though the discovery of many giant viruses along the last decade has changed this view.

Because of their simplicity, viruses cannot replicate outside of the cellular context. They need the cell machinery to make copies of themselves (see chapter 2), and that of course makes a big difference. What is perhaps most impressive of viruses is that they are paramount examples of diversity in all kinds of contexts. They embody a vast range of replication strategies and structural forms of organization, spanning orders of magnitude in genome size and complexity. At the lowest extreme of the complexity continuum are the viroids, which are small folded RNA chains no more than a couple of hundred nucleotides long that do not encode any protein (Flores et al. 2014). At the other end of the complexity are viruses so large that they were initially mistaken for bacteria. This group includes mimiviruses, iridoviruses, pithoviruses, pandoraviruses, and other members of the brotherhood of giant viruses. In figure 1.6 we show a few

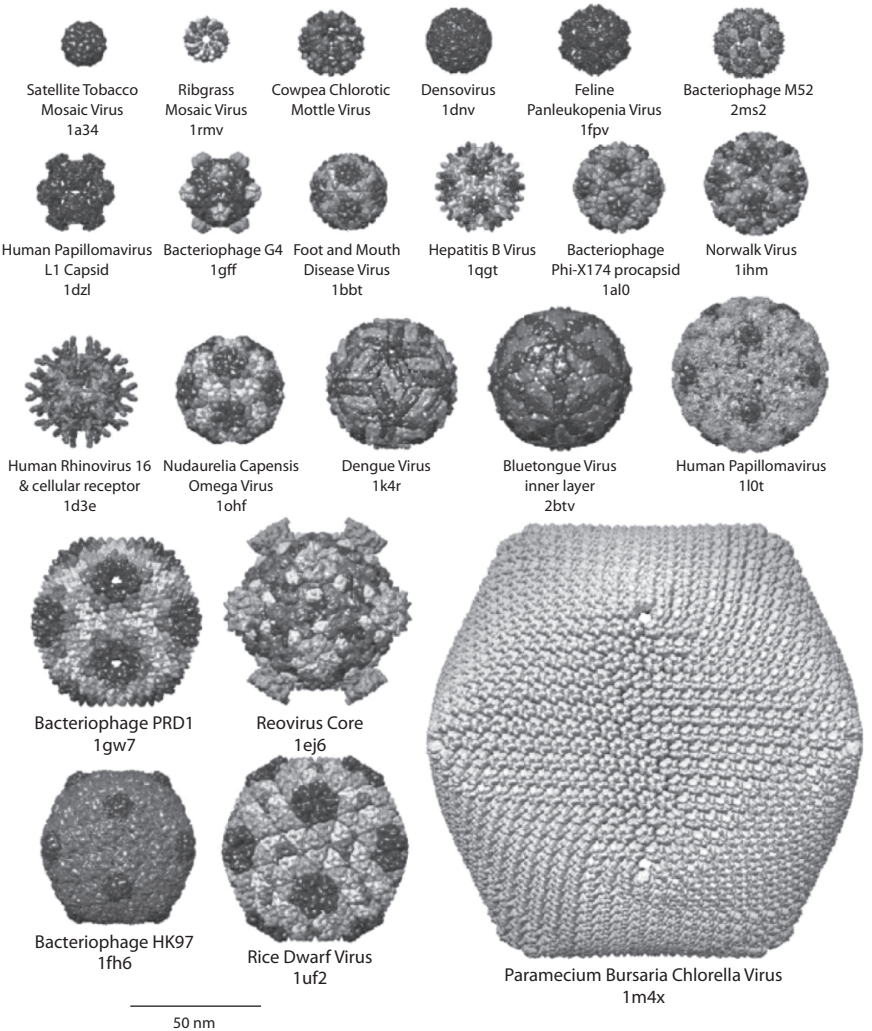

Figure 1.6. Some examples of regular structures found in viruses, from very small, such as $\phi X174$, whose genome was the first to be sequenced, to the largest known mimiviruses, which involve hundreds of genes and have a size even larger than that of the smallest bacteria.

examples of some well-known viruses so we can appreciate the broad range of sizes. The smaller viruses include the famous $\phi X174$, the first entity whose genome (a circular, single-stranded DNA molecule) was fully sequenced (Sanger et al. 1977). The genome of this bacteriophage involves just 5,386 nucleotides, required to encode 11 proteins. But we can also find smaller viruses with some interesting traits besides their tiny size. This is the case of the satellite RNA of TMV, with a 1,063 single-stranded RNA genome which codes just for the capsid and one other protein. This satellite infects tobacco plants already infected with TMV, worsening their symptoms. In this case the satellite virus (that is why this name) needs the cell machinery both of the plant *and* the one from its host virus, TMV.

At the other extreme of the size spectrum, we have a member of the *Mimiviridae* family, including the largest known viruses. The first microscope observations (in 1992) found them infecting amoebas, and given their large size and staining properties they were assumed to be gram-positive bacteria. A correct identification of these microorganisms as true viruses took place eleven years later (La Scola et al. 2003). Since then, many other types have been found (Abergel et al. 2015). Their genome size is comparable with that of cellular genomes, and can be longer than one million base pairs. The finding of this group (as will be discussed in chapter 7) created novel views of the boundaries between living and nonliving entities.

An especially remarkable feature of viruses is their enormous genetic diversity. This diversity is not just a matter of size and composition: it is about the logic of the replication and its evolutionary consequences for the rest of life on earth. There is a striking contrast between the homogeneous nature of information processing that takes place in the nonviral world and what occurs in the virosphere. Cellular genomes replicate thanks to a highly complex molecular machinery based on the transcription of a double-stranded DNA molecule into an RNA chain that

is single-stranded, which itself is then translated by another equally giant molecular complex (the ribosomes) into the proteins necessary to build the whole replication complex (Crick 1970). All cellular organisms respond to this pattern, with very rare deviations. In the virosphere, by contrast, *all* kinds of RNA and DNA combinations and interconversions among them are observable. Such a broad spectrum of genetic strategies allows for a potential evolution that makes viruses a true "genomic laboratory" (Koonin and Dolja 2013). Indeed, one of the first attempts to classify viruses into groups with similar properties was by David Baltimore (1971), based on the type of genetic material (either DNA or RNA, single- or double-stranded) and replication strategy. According to Baltimore's scheme seven groups of viruses can be defined (figure 1.7):

1. Group I is formed by those viruses having a double-stranded (ds) DNA genome. They usually replicate in the nuclei of infected cells and use cellular proteins for their replication. Examples are the herpes viruses and the smallpox virus.
2. Group II includes all viruses having a single-stranded (ss) DNA genome. They also use the cellular machinery for their replication. Examples are the Canine parvovirus and the plant geminiviruses.
3. Group III have dsRNA genomes and replicate in the cytoplasm of the infected cells. They encode for their own replication enzymes. Examples are some fungal viruses.
4. Groups IV and V are the most abundant classes and have genomes of ssRNA of either positive sense (group IV) or negative sense (group V). Positive sense means that the molecule encapsidated can directly be translated by the cellular translation machinery, whereas negative sense means that the molecule

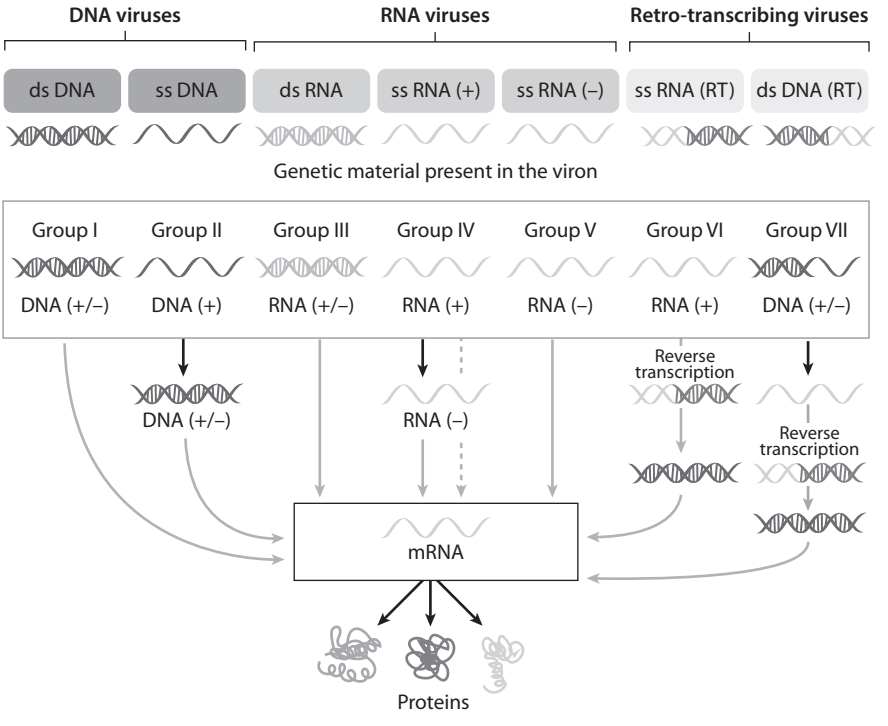

Figure 1.7. A simplified schematic representation of Baltimore's classification of viruses according to the nature of their genomes and their replication intermediates. Adapted from Flint et al. (2015).

encapsidated has to be first transcribed into its complement and then can be translated into proteins by the cellular ribosomes. Most known viruses belong to one of these two families: TMV, *Hepatitis C virus* (HCV), *Foot-and-mouth disease virus*, EBOV, *Yellow fever virus*, and the several influenza viruses.

5. Group VI corresponds to viruses having a positive sense ssRNA genome that is replicated via an intermediate DNA molecule. This group corresponds to the well-known retroviruses whose most characteristic

representative is the HIV-1. All retroviruses encode for an enzyme, the reserve transcriptase, that synthetizes DNA using RNA as template.

6. Group VII corresponds to dsDNA viruses that replicate through an ssRNA intermediate. This small group of viruses, whose representative is *Hepatitis B virus* (HBV), also encodes for a retrotranscriptase (Baltimore 1971; Flint et al. 2015).

No less important here is the fact that viruses have coevolved with their hosts since life began on our planet. This is particularly obvious from the study and sequencing of genomes of plants and vertebrates, which display large amounts of virus-related sequences (Aiewsakun and Katzourakis 2015; Ryan 2016; Mushegian and Elena 2015).

1.4 Viral Planet

Since their discovery, the importance of viruses and the understanding of their ecological and evolutionary impact have only been growing over the last century. It was revealed early that they are responsible for many human diseases, and their genetic plasticity and easy manipulation were crucial in the early days of molecular biology, when bacteriophages (viruses infecting bacteria) were used to test many fundamental ideas concerning the nature of heritable information and the genetic code (Morange 2000; Creager 2002; Cairns et al. 2007). As mentioned above, it was later found that they have a great impact in the ecology of marine ecosystems, with major consequences not only for populations, but for the planet as a whole (see below).

Our biosphere is a complex adaptive system (Levin 1998) where multiple scales of organization are shaped by a number of physical, developmental, ecological, and historical factors. In all these scales viruses play a relevant role. Matter and energy flows

take place through tangled networks of interacting species. A vast range of biomasses are involved, from the largest animals to the smallest cells. But every single organism has at least one, if not many, virus associated to it. Both unicellular and multicellular life forms are rich niches where a virus can find opportunities to evolve. The interactions among hosts and viruses are not always parasitic (see chapter 4) and often lead to disease outbreaks that can have great consequences (discussed in chapter 5). Getting back to the oceans, since viruses have a great impact in the population dynamics of their hosts, and indirect impact on other organisms that prey on their hosts, they also affect deeply the large-scale dynamics of nutrient cycles (Weitz 2016).

Figure 1.8 summarizes the magnitude of the impact of viruses on carbon cycling in the oceans. For comparison, figure 1.8a displays the effects of anthropogenic activities associated to the intensive use of fossil fuels along with the role played by land forests. The basic network of carbon flows[2] is described in figure 1.8b. Viruses kill plankton cells at a rapid pace, leading to both particulate organic carbon (POC) and dissolved organic carbon (DOC). Instead of sinking to the depths, where huge amounts of carbon are being stored, both become suspended or dissolved and thus do not sink. Without the presence of viruses, a large fraction of planktonic carbon is retained in surface waters where it can respire and be photo-oxidized and in chemical equilibrium with the atmosphere. The lytic infection triggered by viruses results in further viral particles and in a complex mixture of molecular pieces forming the cellular debris. This includes small molecules (both monomers and polymers), colloids, and cellular fragments. Most of these components will be incorporated into bacteria and other organisms living in the upper ocean layer (Fuhrman 1999).

[2]Carbon flows are given in Gigatonnes (Gt) per year. A gigatonne indicates one billion tonnes or 10^{15} grams.

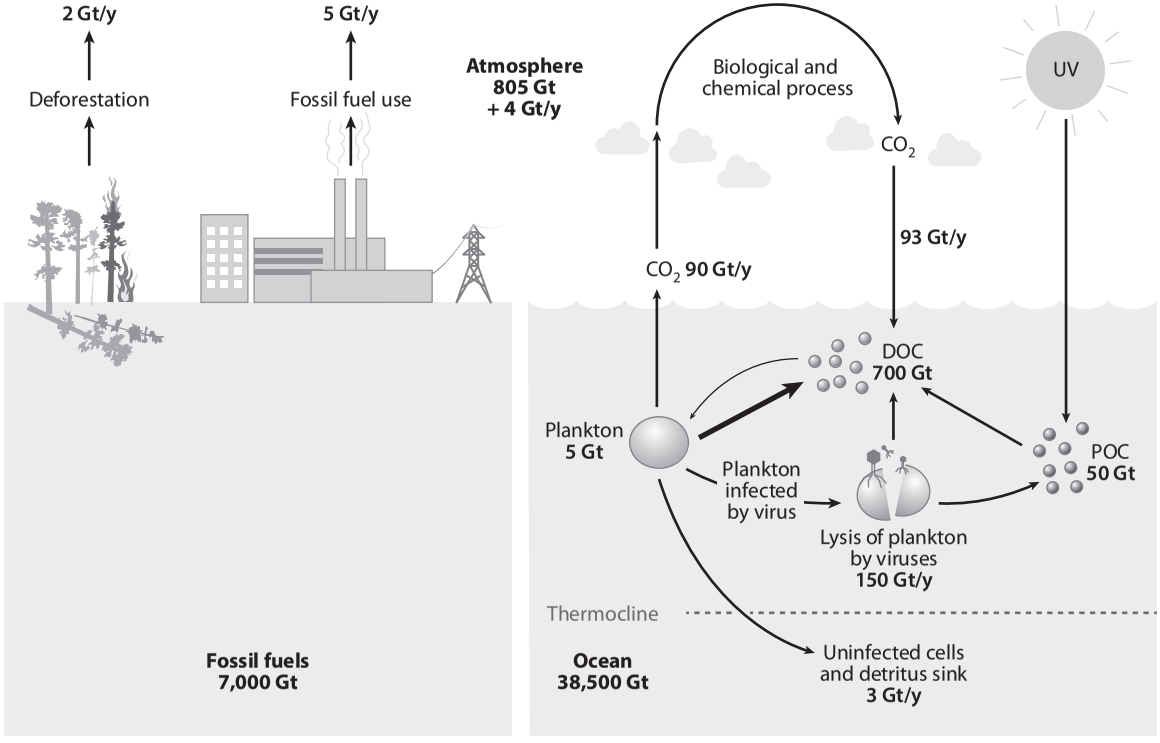

Figure 1.8. Viruses have a deep impact on the carbon cycling at a planetary scale. Because of their effects on the death of plankton, the resulting particulate organic carbon (POC) from cell lysis remains close to the water surface (instead of just sinking into the deep ocean). This strongly affects he balances of CO_2 in the atmosphere.

Viruses define in some ways the coastline of life; little can be understood about evolution of the biosphere without taking viruses into account as major (perhaps dominant) players. It is often said that life is almost everywhere in the planet except inside volcanoes and similar hyper-extreme environments. We can also say that viruses thrive everywhere where life has flourished. This might well be the case for *any* life we find in other worlds, with molecular parasites inevitably associated to self-replicating autonomous entities. In this book we aim to explore the origins of their special status, their universal features and origins from a complex systems perspective. Moreover, viruses are not confined to life. Their properties and propagation dynamics have been an inspiration for understanding the rise of some key evolutionary novelties, such as language and other aspects of cultural and technological evolution.

2

ALIVE OR DEAD?

2.1 Computation and Life

Molecular biology and information technology (IT) emerged almost simultaneously around the middle of the twentieth century and have evolved in parallel since then. Despite the great differences existing between living structures and computer hardware and software, a continuous exchange of ideas and terms took place in the early development of both disciplines (Maynard Smith 2001). An interesting convergence also took place. Engineers building the new technological engines capable of manipulating information used previous theoretical models of computation (as defined by Alan Turing), but they also actively contributed to another important (and much older) domain: the coding and decoding of messages.

Around the 1950s, coding and decoding secret information became a major target of the Cold War efforts. Computer designers and programmers had also to find ways of performing computations at the lowest cost. The early machines were still expensive and had a limited power, and everything needed to be properly designed under strong constraints. That meant writing short, optimized programs, using appropriate coding schemes, and compressing information. A struggle that was being