CHAPTER ONE

What Is a Virus?

1.1 WHAT IS A VIRUS?

Efforts to define a virus inevitably raise the question of exceptions. Nonetheless, a definition or two can help guide us in identifying what is common to all viruses.

> Merriam-Webster's Online Dictionary: an extremely small living thing that causes a disease and that spreads from one person or animal to another.
> Introduction to Modern Virology: submicroscopic, parasitic particles of genetic material contained in a protein coat (Dimmock et al. 2007).

These two definitions are useful, as they reflect the difference in perception as well as current understanding of what a virus is. In that respect, the roots of the term *virus* are also revealing:

> Oxford English Dictionary: late Middle English (denoting the venom of a snake): from Latin, literally "slimy liquid, poison."

Irrespective of source, it would seem that viruses have a bad reputation. Informal surveys tend to yield similar results. For example, when I ask undergraduates to name a virus, some of the most common answers are HIV, influenza, Ebola, chickenpox, herpes, rabies—not a friendly one in the bunch. The answers represent a typical conflation of the disease with the virus. Nonetheless, this conflation is not entirely inappropriate, as viruses do often negatively affect their hosts, whether by causing disease in humans, plants, or animals or killing their microbial hosts.

In fact, one version of the history of viruses begins more or less as follows (Dimmock et al. 2007)—with smallpox. Smallpox is one of the most vicious of diseases, with historical estimates of mortality rates on the order of 30%. Smallpox is caused by a virus, so-called variola, from the Latin *varius* or *varus* meaning "stained" or "mark on the skin," respectively (Riedel 2005).

In 1796, Edward Jenner, a surgeon and scientist, made a bold hypothesis based on the common lore that dairymaids did not suffer from smallpox, perhaps because they had been exposed to an apparently similar disease that affected cows, that is, cowpox. Jenner hypothesized that exposure to cowpox led to protection against smallpox. To test this hypothesis, he transferred material from a fresh cowpox lesion of a dairymaid to an 8-year-old boy who had no prior signs of having been exposed to either cowpox or smallpox. The transfer was likely done with a lancet, directly into the arm of the young boy, who then had a mild reaction—similar to the side effects of modern vaccines—but quickly recovered. Then, Jenner did something remarkable, ghastly, but ultimately providential: two months later he returned and inoculated the same boy with material taken from a new smallpox lesion! Remarkably, the boy did not get sick. This event was widely credited, after Jenner's death, as being the first example of a successful vaccination—as it turns out, a vaccination against what later became known as the smallpox virus.

Viruses as agents of disease and death seem to be the common theme, both in the popular and historical understanding. This bad reputation is similar to that ascribed to bacteria, that is, until recently. Bacteria, which were once considered exclusively "bad" because they cause such diseases as cholera, meningitis, gonorrhea, and chlamydia have had their image redeemed, at least in part. That yogurt companies can market the benefits of products enriched with additional naturally occurring *Lactococcus* cells, that fecal transplants are being considered as a means to stimulate normal digestive tract function, and that the American Society of Microbiology now regularly convenes a meeting on beneficial microbes suggests a reformation in both the scientific and popular opinion of bacteria.

Now, imagine for a moment a yogurt enriched with viruses. This does not seem like a good sales pitch. Or imagine instead, an ocean of viruses. Do you want to go swimming? In fact, a swimmer entering coastal waters for a dip could fill up a single liter bottle and find more than 10 billion, if not 100 billion, virus-like particles. This swimmer is unlikely to get sick, at least not from the viruses. The reasons include the strength of the human immune system and the type of viruses that are found in seawater. Ocean viruses are predominantly viruses of microorganisms and do not have direct effects on human cells. What they do to associated microbes remains an important but ongoing question. Indeed, an alternative history of viruses begins with the viruses of bacteria and constitutes the basis for a far more nuanced view of the range of effects that viruses may have than what is now considered the norm.

This history begins in the late nineteenth/early twentieth century, when microbiologists—also known as "microbe hunters"—such as Louis Pasteur

and Robert Koch were trying to identify the causative agents of disease and to find cures for them (de Kruif 2002). Two microbiologists of the next generation of microbe hunters, Frederick Twort, a British microbiologist, and Felix d'Herelle, a French physician, independently observed a curious phenomenon of clearing in solutions and on plates otherwise replete with bacteria (Twort 1915; d'Herelle 1917). Both Twort and d'Herelle passed the material through a series of filters and chemical preparations that should have eliminated any bacterial or predatory organisms like protists. The filtered material derived from the remains of killed bacteria continued to kill newly grown cultures of cells. Twort thought it was an enzyme that killed bacteria, whereas d'Herelle speculated that a small organism was responsible. He called the small, unseen organism a *bacteriophage* or "bacteria eater," from the Greek word *phagos* meaning "to devour." The notion that viruses could kill bacteria suggested the possibility of phage therapy— the application of viruses to treat human diseases caused by bacterial pathogens. Phage therapy was championed by d'Herelle and became a focus of scientific investigation and a subject of public discourse. Indeed, Dr. Arrowsmith, the protagonist of Sinclair Lewis's *Arrowsmith*, published in 1925, discovers a phage capable of killing the microbe that causes bubonic plague. It would seem that viruses, at least those that infect bacteria, could be forces of good.

Despite these advances, neither Twort nor d'Herelle had seen a virus. This happened later, after the electron microscope was developed and applied to the study of bacteria and viruses in the late 1930s. Mice-associated pox viruses (Von Borries et al. 1938) and the tobacco mosaic virus (Kausche et al. 1939) were two of the first viruses analyzed with an electron microscope. Similar visualizations of bacteriophage followed in 1940 (see discussion in Ackermann 2011). In the debate over whether viruses of bacteria were enzymes or distinct particles, the latter camp ruled the day, helped by the direct observations of viruses. In summary, d'Herelle had been prescient in one significant respect: bacteriophage are a kind of virus—the kind that infects bacteria. But not all the early predictions were realized. Phage therapy is not nearly as commonplace as is the application of antibiotics to treat illness. The reasons for this are treated wonderfully in a book on d'Herelle and the origins of modern molecular biology (Summers 1999). Nonetheless, the promise of phage therapy and phage-enabled therapeutics remains (Sulakvelidze et al. 2001; Merril et al. 2003; Fischetti et al. 2006; Abedon et al. 2011). To the extent that phage therapy can work, it does so because of virus-host dynamics. Similarly, it fails because virus-host dynamics also include evolution (Levin and Bull 2004). Mathematical models that underpin phage therapy models will be introduced and analyzed in Chapters 3–5.

Uncharacteristically for biology, mathematical models were very much part of the formative studies of phage that were designed and executed by luminaries such as Emory Ellis, Max Delbrück, and André Lwoff. Of these, Delbrück was a physicist, and papers from the early days of phage biology (certainly those with his name attached) reveal quantitative thinking that helped build intuition regarding the dynamics that could be seen only at scales far larger than those at which the actual events were unfolding. These early studies provided the foundation for subsequent diversification of the study of phage: the basic concepts of what happens subsequent to infection, experimental protocols for inferring quantitative rates from time-series data, and methods for interpreting and disentangling alternative possibilities underlying the as-yet-unseen actions taking place at micro- and nanoscales (Delbrück 1946; Lwoff 1953). Two recent books revisit these early days, including one written by Summers (1999), mentioned previously, and another by Cairns et al. (2007), *Origins of Molecular Biology*. Both place phage and phage biologists where they deserve to be: at the center of the historical development of molecular biology. These early studies also provided another output: raw material. Phage biologists isolated many of the phage and bacterial strains that have since been disseminated globally for use in many branches of biology (Abedon 2000; Daegelen et al. 2009).

Finally, there is a third story of viruses to tell, one that began only in recent years. It reveals how fraught with difficulties efforts are to define what a virus is and how important it is to think carefully about this seemingly semantic question. In 1992, a French research team identified a previously unknown parasitic organism that infected amoeba. The organism had been observed at least a decade earlier but had not been characterized (for more discussion of this history refer to Wessner (2010)). Each particle was nearly 0.5 μm in size, with a large genome approximately 10^6 bp (base pairs) in length, encapsulated in a membrane vesicle that was itself encapsulated by a protein shell and was further surrounded by fibers. Morphologically the organism had much in common with bacteria, or should have had. As it turned out, this organism is a virus—a *giant* virus. The virus was called a "mimivirus," because of features that suggested it was a *mi*micking *mi*crobe virus (la Scola et al. 2003). Previous research on giant viruses, for example, that by James Van Etten and colleagues, had characterized large *Chlorella* viruses with genomes exceeding 300,000 bp (Van Etten et al. 1991). What made these giant mimiviruses even more remarkable, beyond their size, was that they seemed to constitute a hybrid, chimeric, or seemingly new form of life. Once they infected amoeba cells, these viruses did not depend exclusively on host machinery to produce their component parts. Instead, they carried with them nearly all the genes to do

so themselves. We are now, it seems, at a moment where discoveries call into question the long-standard definition of viruses—these things that live, die, and multiply, just like other organisms.

What, then, is a virus? There are those who say viruses are not alive and others who argue that they are. In the present context, I would prefer to focus attention on ecologically relevant questions, for example, what do viruses do to the hosts and host populations they infect? This question has implications for how entire microbial communities change and function, in part because viruses infect microbial (and metazoan) hosts from the three kingdoms of life (see Figure 1.1). To understand the effect of viruses on microbes and microbial communities it is important to first ask, what are the physical, chemical, and biological dimensions across which viruses differ? These dimensions of viral biodiversity are crucial to understanding viral life history traits and, ultimately, the effects that viruses have on shaping the microbes and the environments in which they persist.

1.2 DIMENSIONS OF VIRAL BIODIVERSITY

1.2.1 Physical

Tobacco mosaic virus, one of the first viruses viewed under an electron microscope, is a rodlike virus approximately 300 nm in length and 20 nm across. In contrast, phage λ, a subject of formative studies of gene regulation (Ptashne 2004), has a capsid approximately 50 nm in diameter with a tail fiber extending approximately 150 nm. Although viruses are "small," their range of sizes is larger than is widely recognized, spanning at least one, if not two, orders of magnitude, with significant morphological variation when considering viruses that infect all kingdoms of life (Figure 1.1). That size varies by two orders of magnitude is due to the recent discovery of "giant" viruses that can reach 0.5 μm in size that infect amoeba, ciliates, and perhaps other eukaryotes (Van Etten et al. 2010).

The physical disparity in size might seem curious, or simply a curiosity (akin to the "Rodents of Unusual Size" from *The Princess Bride*). However, differences in the physical size of virus particles are linked to many aspects of viral life-history traits. The study of the relationship between size and function is one of the oldest in science. Leonardo da Vinci is considered the first to develop an argument for allometric scaling in biology (see discussion in Brown and West (2000)). Da Vinci hypothesized that the sum of cross-sectional areas of tree limbs should be equal before and after branching and,

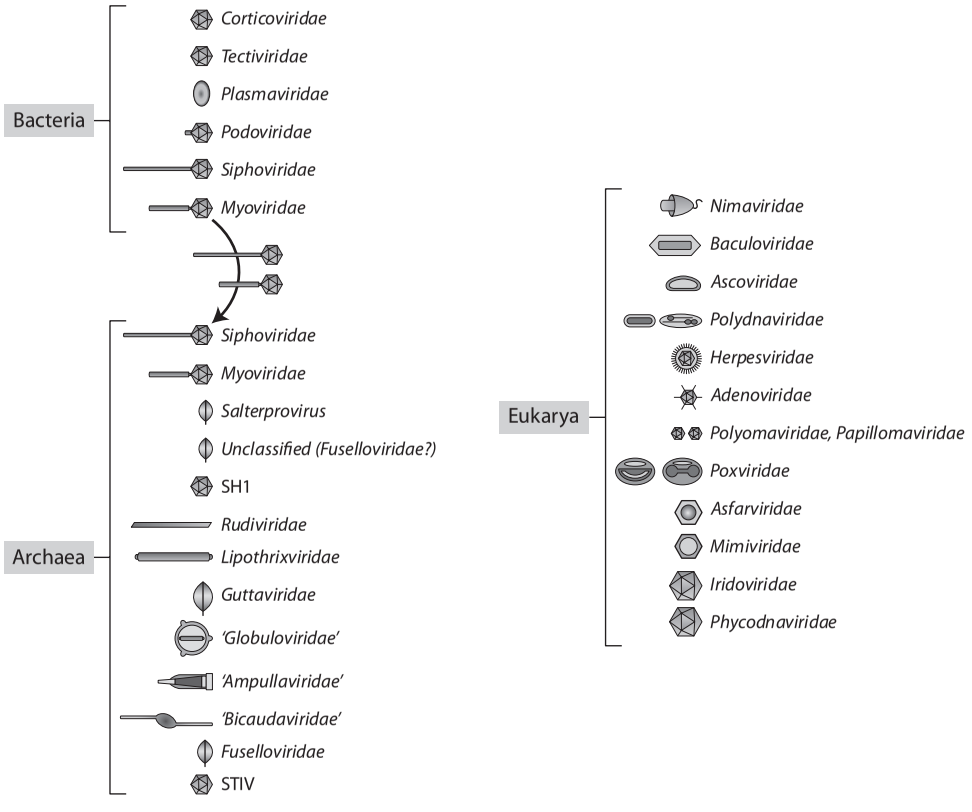


FIGURE 1.1. Morphological diversity of viruses that infect a diversity of microbial and metazoan hosts. Image redrawn based on the original image in Figure 6 of the review by Prangishvili et al. (2006). The original caption reads: "The virus families listed are approved by the International Committee on Taxonomy of viruses, and the schematic representation of virions (not drawn to scale) are presented as in Fauquet et al. (2005). Proposed families are shown in inverted commas. SH1, *Haloarcula hispanica* virus 1; STIV, *Sulfolobus turreted* icosahedral virus."

further, that limb lengths scale with limb diameter. This isometric change in limb component size is not quite accurate, given that limb widths increase in size faster than do limb lengths (McMahon 1973). The modern history of the study of *allometry*, the change of organismal structure and function with body size, has its origins in the late 1800s. In 1883, the German physiologist Max Rübner claimed that the metabolic rate of dogs could be estimated accurately based on knowledge of their size alone (see discussion in Kleiber (1947)). The reasoning assumed that an organism's metabolic rate was mediated via exchange with the surroundings. Organismal surfaces were hypothesized to scale with body size to the 2/3 power, scaling in some sense like spheres.[1] If exchange area scaled to the 2/3 power, then so, too, would metabolic rate. This simple hypothesis is not that far off, though how far off such a prediction is depends on whether mice or elephants are being considered. The analysis of the scaling of metabolic rate is a matter of long-standing scrutiny and debate (Kleiber 1961; McMahon 1973; Peters 1983; Schmidt-Nielsen 1984; Brown and West 2000). Indeed, linking organismal body size to organismal function, such as rates of locomotion, predation, and even death, is the basis for the study of macroecology (Brown 1995).

What is the analogous link between size and function in the case of virus–microbe interactions? Here, there are two sizes to consider: the size of the virus and the size of its host. This chapter largely focuses on virus size, which has two key components: the size of the virus particle and the length of the virus genome. These two sizes are interrelated. Viral genomes are packed under pressure inside a protein capsid. In the case of dsDNA nonlipid-containing phage, the genome is highly organized; for example, there is evidence that DNA can be coiled (Purohit et al. 2005) or even folded toroidally (Petrov and Harvey 2007). The total volume of the genome can be approximated as the sum of the volumes of the nucleotides. The available volume inside the capsid is $4\pi r^3/3$, where r is the internal radius of the capsid. How does the realized volume of the genome change as the available volume increases?

Figure 1.2 shows the measured empirical relationship between the number of base pairs and capsid internal radius, r. The relationship is linear on a log-log plot, with an exponent of 3 and a prefactor of 2. The prefactor is not universal, and requires use of nanometers as units for length. In other words,

$$n_{bp} = 2r^3 \tag{1.1}$$

[1] An old joke about physicists and spherical cows comes to mind; here it seems to apply to physiologists and spherical dogs. Nonetheless, such spheres are not a bad starting point.

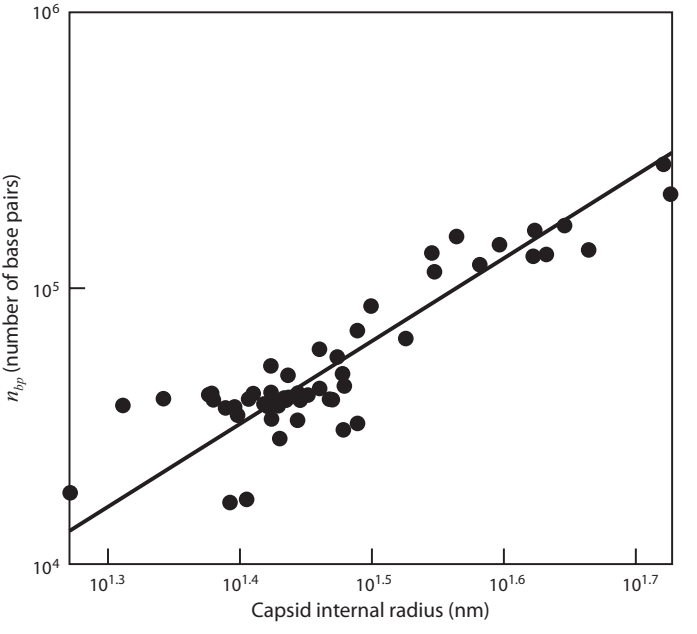


FIGURE 1.2. The scaling of genome length measured in number of base pairs versus the internal radius of the capsid. Quantitative relationship between genome length and capsid size for nonlipid-containing dsDNA bacteriophage for 54 phage types (see Jover et al. (2014) for full accession information). The solid line corresponds to the best-fitting cubic; that is, $n_{bp} = cr^3$, where $c \approx 2.0 \pm 0.2$. Reprinted from Jover et al. (2014).

for icosahedral capsids. Hence, a virus with an internal radius of 25 nm is predicted to have approximately 31,250 bp, and a virus with an internal radius of 50 nm is predicted to have approximately 250,000 bp. This relationship between viral size and genome length has been tested against nonlipid-containing dsDNA phage (Jover et al. 2014). The relationship is a useful baseline for quantifying other life history traits, as explained in the next section. The empirical relationship can also be used to derive the *fill* of a capsid, that is, the fraction of volume within a capsid taken up by the phage genome:

$$n_{bp} = \text{fill}\frac{v_{ic}}{v_{bp}} = \text{fill}\frac{4\pi}{3v_{bp}}(r_c - h)^3, \tag{1.2}$$

where v_{ic} is the volume inside the capsid, r_c is the outer radius of the capsid, h is the thickness of the capsid, and v_{bp} is the volume of a base pair. Noting that $r = r_c - h$, and given the constants $h \approx 2.5$ nm and $v_{bp} \approx 1.07\,\text{nm}^3$, one can

infer that fill $= 0.51 \pm 0.04$ (Jover et al. 2014). In summary, scaling analysis reveals that phage genomes take up approximately 50% of the available volume inside the capsid, irrespective of whether the capsid is small or large. Similar scaling arguments can be used to estimate the number of proteins making up the capsid, n_{pr}. The core notion is that the capsid can be represented as a spherical shell with volume v_c and uniform thickness h. The expected number of proteins in the capsid is

$$n_{pr} = \frac{v_c}{v_{pr}} = \frac{4\pi}{3v_{pr}}\left[r_c^3 - (r_c - h)^3\right] \tag{1.3}$$

$$= \frac{4\pi}{3v_{pr}}(3r_c^2 h - 3h^2 r_c + h^3). \tag{1.4}$$

The number of base pairs increases, to leading order, as r_c^3, whereas the number of proteins increases, to leading order, as r_c^2. Such information is also key to estimating the elemental content of virus particles.

Viruses of microbes vary in other ways as well. As should be apparent from Figure 1.1, viruses differ not just in size but also in shape. This is true whether considering the viruses of bacteria, of archaea, or of microeukaryotes. The study of environmental phage isolates often begins with the question, is the isolate a myo, a sipho, or a podo? This lingo stands for myoviruses, siphoviruses, and podoviruses. The question means, is the virus from the family Myoviridae, Siphoviridae, or Podoviridae, respectively. All three are tailed viruses but are distinguished most readily by their tails. Myoviruses have long, contractile tails; siphoviruses have long, noncontractile, flexible tails; and podoviruses have short tails (Figure 1.3). Perhaps the best known representatives of these three families are T4, phage λ, and T7, which have tail lengths of approximately 140 nm (Kostyuchenko et al. 2005), 150 nm (Katsura and Hendrix 1984), and 20 nm (Krüger and Schroeder 1981), respectively. By comparison, environmental assays of tail lengths of marine viruses can be used to infer that lengths are typically 150 nm for myoviruses, 210 nm for siphoviruses, and 15 nm for podoviruses (Brum et al. 2013)—though many viruses are non-tailed. The tails of viruses have functional roles, particularly in defining host specificity and infection initiation (e.g., see the detailed analysis of entry in the case of T7 in Hu et al. (2013).

1.2.2 Chemical

Virus particles can comprise a head and, sometimes, a tail. The head is a protein capsid surrounding genetic material, either RNA or DNA, whereas the tail is made up of proteins. The protein capsid or "coat" may include various

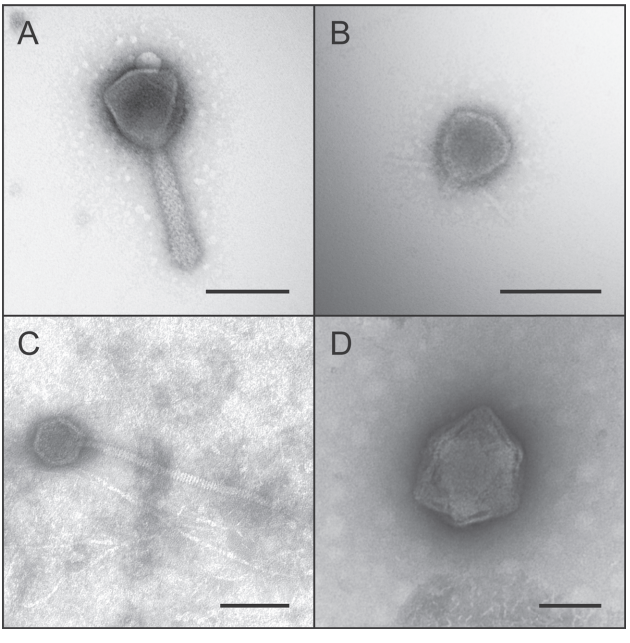


FIGURE 1.3. Electron micrographs of virus particles from ocean-surface samples: (a) myovirus; (b) podovirus; (c) siphovirus; (d) nontailed virus. The samples were negatively stained. Scale bar denotes 100 μm. Images are reproduced with permission of the copyright holder, Jennifer Brum.

decorations. Structural virologists have developed elegant systems for describing the types of folds and configurations observed in capsid assembly (Jiang et al. 2003). The structure of virus particles may itself be informative with respect to the evolutionary history and origins of viruses (Bamford et al. 2005). Yet, the protein stoichiometry of viral capsids may be less important than the elemental stoichiometry of virus particles, from the perspective of ecology and ecosystem functioning. Why is this the case? In natural environments with scarce resources, there are limited opportunities to ingest and digest a particle that is rich in both phosphorus (P) and nitrogen (N). Hence, a virus may be a tasty snack to certain nanoflagellates (Gonzalez and Suttle 1993), because viruses contain a relatively high proportion of nitrogen and phosphorus per unit mass, compared with alternative food sources (like bacteria or other phytoplankton). How often this snack is ecologically relevant depends on the availability of other potential prey items and the ability of the consumer to utilize a virus as food. The snack also involves a certain risk: taking up a virus may lead to infection.

How much carbon (C), nitrogen, and phosphorus is in a virus particle? The capsid is made up of proteins and has an external radius r_c. The external radius

is the distance from the center to the outer boundary of the capsid. The number of base pairs inside the capsid, n_{bp}, scales as $(r_c - h)^3$, where h is the thickness of the capsid (Eq. 1.4). This scaling presumes that the genome occupies a fixed fraction of the available volume. In contrast, the expected number of proteins in the capsid, n_{pr}, scales as r_c^2. This means that the relative number of proteins to base pairs decreases with increasing viral size. Biochemically, this relationship has an important consequence, as nucleotides and proteins have distinct molecular compositions.

The average molecular formula of a base pair (i.e., a pair of nucleotides), expressed in terms of C:N:P is 19.5:7.5:2. That is, there are 19.5 molecules of carbon for 7.5 molecules of nitrogen for 2 molecules of phosphorus in every base pair. The "average" here assumes an equal probability of having an A:T base pair as a G:C base pair, whose molecular compositions are distinct. Deviations are expected for any given genome sequence. Nonetheless, assuming 50% GC content provides an important baseline for assessing the elemental composition of viruses. In contrast, the amino acids that constitute proteins have no phosphorus, but they do contain carbon and nitrogen—again, the particular ratio depends on amino acid composition. Analysis of primary sequence information for more than 2000 viral proteins reveals that they have, on average, 31 molecules of carbon and 8.7 molecules of nitrogen/nm^3 of protein (Jover et al. 2014); that is, a C:N ratio of 3.6:1—slightly more carbon rich and nitrogen poor than DNA.

Jover et al. (2014) demonstrated how to combine the scaling of molecular composition at the level of nucleotides and proteins with the elemental composition of such molecules to arrive at a predictive model for the elemental composition of virus heads. The combination involves the addition of the elemental composition of the two components of a virus head, its genome and its capsid:

$$\mathrm{C}_{\text{virus head}} = \mathrm{C}_{\text{genome}} + \mathrm{C}_{\text{capsid}}, \tag{1.5}$$

$$\mathrm{N}_{\text{virus head}} = \mathrm{N}_{\text{genome}} + \mathrm{N}_{\text{capsid}}, \tag{1.6}$$

$$\mathrm{P}_{\text{virus head}} = \mathrm{P}_{\text{genome}}, \tag{1.7}$$

After the size-scaling relationship and the chemical composition of molecules are combined, the specific predictions are

$$\mathrm{C}_{\text{virus head}} = 39(r_c - 2.5)^3 + 130(7.5r_c^2 + 18.75r_c + 15.63), \tag{1.8}$$

$$\mathrm{N}_{\text{virus head}} = 15(r_c - 2.5)^3 + 36(7.5r_c^2 + 18.75r_c + 15.63), \tag{1.9}$$

$$\mathrm{P}_{\text{virus head}} = 4(r_c - 2.5)^3, \tag{1.10}$$

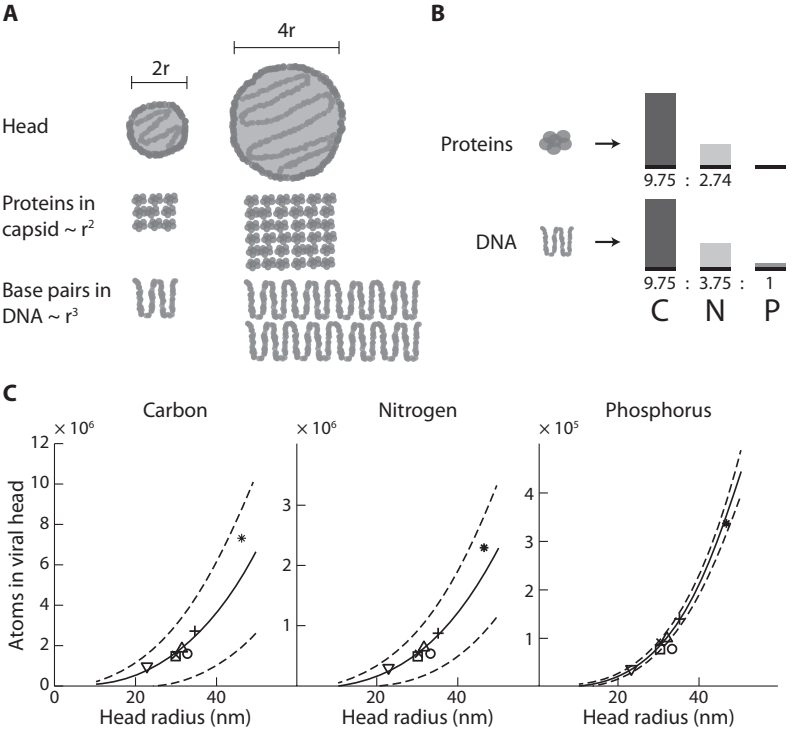


FIGURE 1.4. C, N, and P content of the viral head as a function of its external radius (solid lines). The data correspond to experimentally obtained contents of C, N, and P for different viral heads: ∗ T4, + N4, × Syn5, △ λ, ○ HK97, □ T7, ▽ ϕ29.Full information on data sources can be found in Jover et al. (2014). Reprinted from Jover et al. (2014).

where r_c is in units of nanometers. Predictions for the number of atoms of carbon, nitrogen, and phosphorus were then compared against the total elemental content as enumerated for seven viruses: T4, N4, Syn5, λ, HK97, T7, and ϕ29. The criterion used in selecting these viruses was that their genome sequence and complete capsid structure was available. The latter requirement proved more restrictive. Many viral genomes have been sequenced, but very few entire 3D models of dsDNA viruses of microorganisms are available in which the corresponding amino acid sequences of each protein are known. Model predictions, and variation due to uncertainty in model parameters, are shown in Figure 1.4. As should be evident, the model captures the size dependence of elemental content in actual virus heads, yet the formalism can easily be extended to virus tails. Because tail proteins do not contain phosphorus, the total P content is bound in the head only, whereas the C and N inside a virus

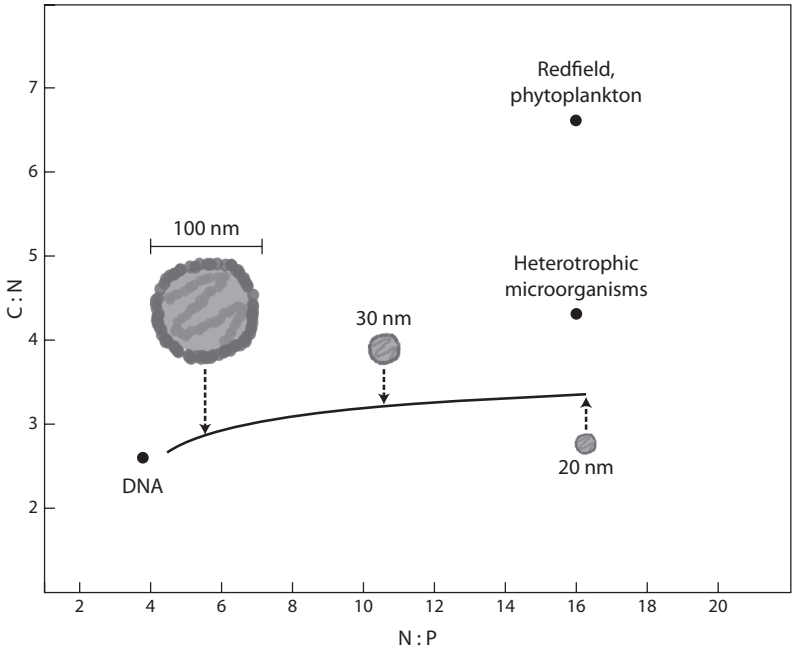


FIGURE 1.5. Size dependence of elemental stoichiometry for viral heads. The *x*-axis is the N:P ratio, and the *y*-axis is the C:N ratio. The three reference points plotted correspond to DNA (given a G:C ratio of 0.5), the Redfield ratio of 106:16 (for C:N) and 16:1 (for N:P), and an approximate consensus for heterotrophic bacteria. Reprinted from Jover et al. (2014).

head underestimates the total C and N inside a virus particle by approximately 10%–15% for this group of myoviruses and siphoviruses (Jover et al. 2014).

Analyzing the elemental content of viral heads leads to an important conclusion: the elemental stoichiometry of virus particles differs from that of microbial hosts. The term *elemental stoichiometry* refers to the ratio of atoms of distinct elements inside an organism (Sterner and Elser 2002). Viruses, owing to their high concentration of DNA, are predicted to have C:N ratios that remain relatively invariant. This relative invariance reflects the similar C:N ratios of proteins and nucleotides—the building blocks of most virus particles with the exception of those that contain lipids and carbohydrates. In contrast, the N:P ratio of virus particles is predicted to range from approximately 16 to 4. Together, these ratios imply that there may be as few as 15 atoms of carbon for every atom of phosphorous in a virus particle. This composition is remarkably nutrient rich, given that there are usually greater than 50, if not greater than 100, molecules of carbon for every molecule of phosphorus in a microbial host cell (Figure 1.5). The differential scaling of carbon, nitrogen, and phosphorus

in a virus compared to that in its host has a number of implications for marine biogeochemical dynamics—issues that will be revisited in Chapters 6 and 7.

Infection of host cells provides another avenue for exploring chemical diversity associated with viruses. Virus infection of host cells redirects host cell metabolism toward production of viruses. Yet, in some virus-host systems, the structural elements of virus particles include novel macromolecules that are produced only in the course of virus infections. A prominent ecological example is the interaction of viruses with the algal phytoplankton host *Emiliana huxleyi*—the most abundant coccolithophore in the global oceans, and a key population in driving carbon cycling (Iglesias-Rodríguez et al. 2002). When infected by the algal virus EhV86, host cells undergo a dramatic change in chemical profile. Of note, the profile of lipids changes from host-associated lipids to virally encoded glycosphingolipids (Bidle and Vardi 2011). The molecular weight of virally encoded lipids reaches levels of up to 200 fg per cell (similar to the total mass of an *E. coli* bacterium). Hence, the chemical diversity associated with viruses can represent novel biomarkers for identifying infection as well as mediating chemical arms races in natural systems. Underlying such chemical diversity is the biological diversity encoded in viral and host genomes.

1.2.3 Biological

It has been said—repeatedly—that viruses infect all manner of life. This makes for a good initial hypothesis. But the statement itself is vague and fundamentally unanswerable. Alternatively, one could ask, are all living organisms *potentially* infectable by a virus? Or are all organisms currently infected by a virus? Or perhaps, will all organisms, at some point, become infected by a virus? One could also ask, which *types* of organisms can viruses infect? Are there viruses that infect individual microorganisms from the three domains of life: Archaea, Bacteria, and Eukaryota? The answer to this last question is yes for Archaea (Prangishvili et al. 2006; Prangishvili 2013), yes for Bacteria (Calendar 2005), and yes for Eukaryota, including microbial eukaryotes (Van Etten et al. 1991, 2010). Because there is significant diversity within these three kingdoms of life, it is unknown, how many types of viruses infect any given host sampled from the environment. A follow-up question of relevance is, do viruses infect each type of archaea, bacteria, or eukaryote at finer taxonomic scales, for example, family, genus, or even species?

This question—currently and likely permanently unanswerable—serves as a useful *Gedankenexperiment*, "thought experiment." The reason why the problem is intractable has something to do with the nature of viruses and

the nature of the diversity of their microbial hosts. The gold standard for viral identification is *isolation*, that is, the separation of a particular virus from the rest of the community. To isolate a virus, one must first isolate a host upon which the virus can replicate. This precondition leads to many potential biases in what is known about the ecology of viruses and their microbial hosts, because current information is largely based on the few examples of viral isolates, which requires isolating bacteria, archaea, and/or microeukaryotes. Such isolation is difficult (Rappé and Giovannoni 2003). There are many more microbial individuals in the environment than can be isolated as yet in the laboratory. This discrepancy is known as the "great plate count anomaly" (Staley and Konopka 1985). The anomaly refers to the large difference between estimates of microbial abundances using culture-based techniques and culture-independent techniques. Estimates of the number of bacteria in seawater from culture assays typically yield 10^4/ml, when grown on seawater-based sterile media. A culture-independent approach to stain the sample so that particles that have DNA and are approximately 0.5–2 μm in cross section typically yields estimates of microbial abundances of 10^6/ml. As a consequence, culture-based estimates often undercount the true abundances of microbes by two orders of magnitude (Rappé and Giovannoni 2003). The reasons for this paradox are many, yet can be summarized as follows: the conditions necessary to cultivate microbial populations are not yet known (Leadbetter 2003). These conditions may even include a requirement to live with, or at least among, other organisms—suggesting fundamental limits to current culture-based efforts at isolation.

The diversity of viruses extends to, and indeed is encoded in, viral genomes. The genome structure of viruses differ from that of hosts—in similarity to one another, size, and compactness. Comparisons of viral genomes are made more difficult given that they do not possess universal marker genes, like those that encode 16S rRNA within microbes and metazoans (Pace 1997). Instead, the differences among viruses can be compared at coarse and fine scales. Coarse-scale comparisons are facilitated by comparing the genomes or imputed proteomes of viruses and using these distances as the basis for constructing trees (Rohwer and Edwards 2002). Fine-scale comparisons among viruses focus on microevolutionary changes in viral genomes, such as may occur during adaptation of influenza viruses (Koelle et al. 2006) to bacteriophage λ (Meyer et al. 2012). In either case, variations in viral-encoded genes help determine which hosts a virus can infect and what happens to those hosts after they are infected.

Viruses face pressure to reproduce rapidly while evading host defenses and are thought to have evolved a highly compact genome. One measure of this

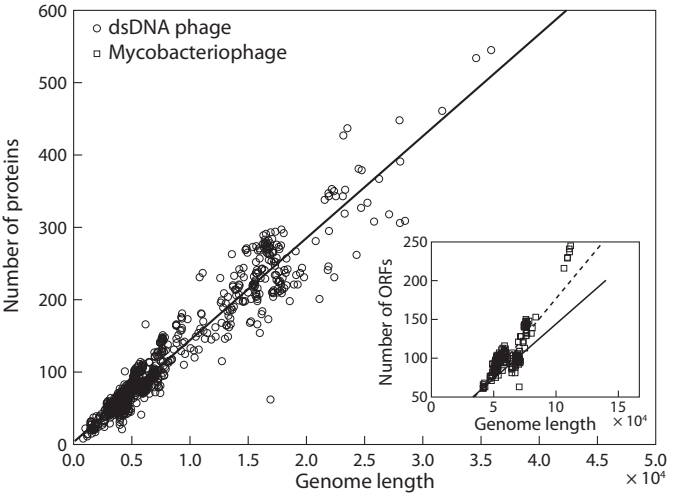
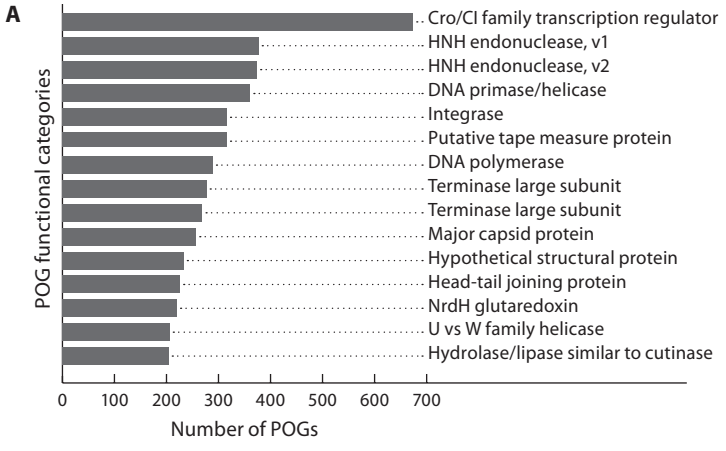


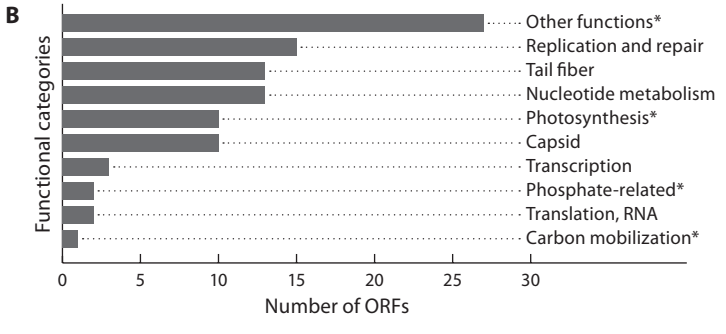
FIGURE 1.6. The number of putative proteins as a function of genome size. (Main) Relationship as estimated from 1124 sequenced dsDNA bacteriophage genomes. The solid line is the line of best fit (slope of 1.41×10^{-3}, corresponding to one open reading frame (ORF) for every 709 nucleotides, assuming no gene overlaps). For this fit, $R^2 = 0.93$. Genome sequences and annotations were downloaded from the National Center for Biotechnology Information (NCBI) in March 2014. (Inset) Relationship as estimated for mycobacteriophage. Each square in the inset represents one of 253 sequenced phage genomes designated as a siphovirus for which an ORF annotation had been completed. The dashed line is the line of best fit (slope of 1.94×10^{-3}, corresponding to one ORF for every 515 nucleotides, assuming no gene overlaps). For this fit, $R^2 = 0.74$.

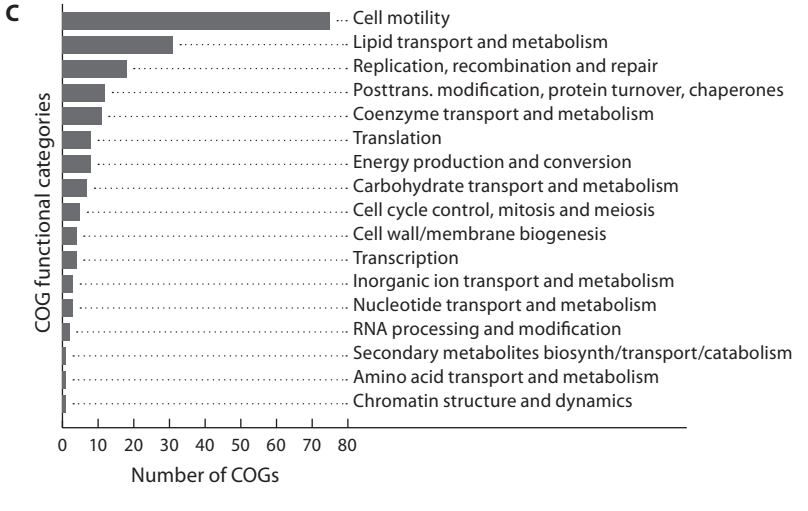
compactness is evident in Figure 1.6, in which the number of proteins is compared against genome length for dsDNA phage (see Kristensen et al. (2013)). The data are well fit by a line whose slope implies that viral genes are approximately 709 nucleotides in length. The linear fit explains 93% of the variation. Another way to interpret this result is that given a broad sample of viruses of differing sizes, the number of genes can be estimated by dividing the genome length by 709. In fact, viral genes can overlap, suggesting viral genomes are even more compact than a strict division would suggest. Figure 1.6 inset shows further evidence that viral genomes are compact. The analysis focuses on 253 siphoviruses from the well-studied mycobacteriophage clade that have been isolated as part of the HHMI Phage Hunters effort (Hatfull et al. 2008; Hatfull 2012). Increases in viral genome length correspond to a proportionate increase on average in the number of genes, unlike human genomes, which are largely noncoding. Given viral genomes that vary from a few kilobases to hundreds of thousands of bases, what are those genes and what do they do?

Estimates of the total number of phage-encoded genes presents one view of the scope of the problem. More than a decade ago, Forest Rohwer suggested that there are likely more than 2 billion genes encoded among phage on Earth, the vast majority of which have not been discovered (Rohwer 2003). In 2013, Matthew Sullivan and colleagues performed a similar analysis based on clustering of viral proteins, revising estimates to a range of 0.6 million to 6 million genes encoded in the global phage virome (Ignacio-Espinoza et al. 2013). This discrepancy between estimates of viral gene diversity is due, in part, to the use of different thresholds by which two genes are considered part of the same or different groups. Moreover, both estimates rely on inference methods meant to extrapolate the diversity of a community from that of the sample. Then, and now, the global phage virome was and is deeply undersampled, which poses a problem to estimation based on extrapolation. For now, it is sufficient to point out that the total number of phage genes, and indeed of viral genes, is large—certainly larger than a few million. Definitive estimates of the phage gene pool size require careful consideration of the problems inherent in such deep extrapolations. This point will be developed further in Chapter 6.

Irrespective of the total size of the viral gene pool, understanding the function of putative viral genes is also difficult. Such understanding depends, in some sense, on being an expert first, that is, identifying functions for those open reading frames that code for putative proteins. The diversity of viral genes is immense compared with the diversity corresponding to a relatively few well-studied model systems. As a consequence, databases of documented viral functions remain sparsely annotated. For those unfamiliar with what makes up a virus, it is worthwhile to revisit the "usual suspects" of functions that are commonly found in viral genomes, and for which a function can be hypothesized. It is not yet practical to analyze the genomes of known viruses one at a time, but thankfully, Eugene Koonin's group has been thinking about related problems for a long time (Koonin et al. 2002; Koonin and Wolf 2009). By analyzing thousands of bacteriophage genomes, David Kristensen, Eugene Koonin, and collaborators have proposed a system for categorizing putative viral proteins into *phage orthologous groups*, or POGs (Kristensen et al. 2011, 2013). POGs denote viral genes that are thought to have similar functional roles and common evolutionary origins. POGs should be thought of as analogous to *clusters of orthologous groups*, or COGs. The concept of COGs has proved instrumental in categorizing genes found in diverse organismal types in terms of similar function (Tatusov et al. 2000). When these clustering approaches are applied, the "top" categories of POGs include some genes coding for functions that should be familiar to virologists. These functions







include structural proteins, enzymes that help viruses integrate into their host genomes, enzymes that help viruses replicate, and regulatory proteins that shape the fate of the infected cell. The relative frequency of appearance of the top 20 POGs compiled by Kristensen et al (Kristensen et al. 2013) can be seen in Figure 1.7. This list represents a starting point, particularly for those outside the field of viral biology, for recognizing some of the most common types of genes inside viral genomes.

What is less often appreciated is the extent to which viruses encode genes that encode for a diversity of functions. Consider the myovirus P-SSM2, which infects ubiquitous cyanobacteria of the genus *Prochlorococcus* (Sullivan et al. 2005). This cyanophage genome, and others closely related to it, includes a number of "unusual" genes, such as *psbA*, *psbD* and *pstS* (Figure 1.7). *psbA* and *psbD* are phage-encoded photosystem II genes. These genes are expressed during infection and augment the production of photosynthetic machinery, which is then redirected toward the viral pathway. Importantly, these genes were not "stolen" just recently from host cells—phylogenetic comparison of gene sequences suggests evolution inside both phage and host genomes, as well as mixing between phage and hosts (Sullivan et al. 2006). Similarly, *pstS* is a phosphate-inducible gene that has been hypothesized to increase the uptake of phosphorus by infected cells, a relevant factor in low-nutrient environments. In summary: viruses encode genes that don't pertain "just" to structural components, integration, and escape. Viruses also encode genes that modify metabolic pathways during the infection cycle.

Another example of unusual genes encoded in viral genomes derives from the mimivirus, a so-called giant virus that infects amoeba (la Scola et al. 2003). The mimivirus genome is nearly 1.2 Mbp in length with an estimated 1184 genes (Raoult et al. 2004). The mimivirus genome includes many genes that are highly unusual for viruses, though seemingly common for microbes. Functional categories of these genes include associations with cell motility,

FIGURE 1.7. Viral gene diversity. (A) Functions of the top 20 phage orthologous groups (POGs), ranked in terms of the number of viral genome isolates in which they appear. Adapted from Figure 2 of Kristensen et al. (2013). The most highly represented POGs correspond to the usual suspects of phage genome composition. Nonetheless, they do not reflect all the variety in viral gene function, as illustrated by the presence of photosystem and nutrient-inducible genes found in cyanophage (B) and many cellular-analogue genes found in mimiviruses (C). (B) Functional categorization of cyanophage, including common and "unique features" marked with *-s. Adapted from Table 5 of Sullivan et al. (2005)). (C) Functional categorization of mimivirus based on clusters of orthologous genes (COGs). Adapted from Table 3 of Raoult et al. (2004).

energy production, lipid transport, lipid metabolism, cell wall/membrane biogenesis, and even chromatin structure (Figure 1.7). The steps of a mimivirus infection unfold quite unlike those following phage infection of a bacteria, in which host cell machinery is responsible for the bulk of transcription and translation. In contrast, during a mimivirus infection the site of transcript production resides in a localized structure called the "virus factory" (Suzan-Monti et al. 2006)—similar to that in virus infection cycles of other amoebae. The mimivirus and other giant viruses harbor genes to generate new virus particles that emerge from the viral factory. The infection cycle includes many of the steps that would otherwise be associated with a typical cell cycle, with many steps still the topic of active research.

These two examples serve to illustrate a larger point: the diversity of functions in viral genomes annotated in sequence databases reflects the ability of researchers to isolate viruses. Isolated viruses are not a random sample of the environment. Viruses are highly diverse, in that they can persist in hot springs (Held and Whitaker 2009; Snyder and Young 2011), soils (Williamson et al. 2007), lakes (Heldal and Bratbak 1991), oceans (Breitbart 2012), and inside microbiomes (Minot et al. 2013). Adaptation to hosts in such varied environments is associated with concomitant genetic diversity. These discoveries are ongoing. Indeed, in 2014 a French team of researchers reported the discovery of giant viruses related to the mimivirus family. These viruses were frozen in the Siberian tundra for more than 30,000 years, were revived, and still retained the ability to infect extant amoeba cells (Legendre et al. 2014). There is evidently more to discover on all three components of viral diversity outlined in this chapter. To understand the ecological role of viruses requires moving outward, from virus particles to virus–host interactions. That is the subject of Chapter 2.

1.3 SUMMARY

- The study of virology is relatively recent. The conclusive discovery that viruses were a causative agent of infection in plants, animals, and microbes was not made until the advent of electron microscopy in the 1930s.
- Viruses vary in size from genomes of a few thousand to more than a million nucleotides.
- Viral capsids vary in linear dimensions from approximately 20 nm to more than 400 nm in diameter.
- The elemental composition of viruses can be predicted based on simple scaling arguments.

- Virus particles are relatively nutrient rich compared with their hosts.
- Viral genomes are compact, with the number of putative genes scaling linearly with genome size.
- The functional diversity of viruses includes many canonical viral genes, such as those that code for capsid proteins and transcriptional regulators.
- The functional diversity of viruses includes many noncanonical genes, such as those that code for proteins that are part of photosystem pathways or cell-wall pathways.