Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Getting Ready for the Era of Comparative Genomics: The Importance of Viruses

Virologists like to recollect the episodes in the history of science in which viruses played an important role. Max Delbruck and the Phage Group at Cold Spring Harbor Laboratory is the most famous example (“Phage and the Origins of Molecular Biology,” 2006). Many other fundamental discoveries in molecular biology were facilitated by the simplicity of viral genetic systems, with their limited number of tractable molecular components. This includes the Hershey and Chase experiment on phage T4 infectivity, which settled the question of whether genes were made of protein or nucleic acid (Hershey and Chase, 1952); Fraenkel-Conrat’s demonstration of infectivity of TMV RNA (Fraenkel-Conrat et al., 1957); virus self-assembly studies, also by Fraenkel-Conrat as well as others (reviewed in Fraenkel-Conrat, 1990); understanding of retrovirus genome strategy, which led to refinement and crystallization of the central dogma of molecular biology (Crick, 1958, 1970); and the discovery of RNA splicing (Berget et al., 1977; Chow et al., 1977).

These occurred decades ago. But how about now? Today, we have bacteria, yeasts, the nematode Caenorhabditis elegans (which has no common name, although “elegant worm” seems to be gaining popularity), flies, Arabidopsis thaliana (too many common names, including mustard weed, Thale’s cress, and mouse ear, none of them gaining much traction), sea squirts, mice, mosses, and so on. All of these are immensely useful model systems, some simpler than others, but all with the number of genes in the thousands. The simplicity of genome does not seem to be a major requirement for a model system anymore. So what about viruses—are they still of any use as models of anything? In a world full of interesting living species, who cares about viruses, except for virologists?

There is no doubt that the medical, agricultural, and other social impact of viruses is on the scale from moderate to huge, depending on the disease. But more important to the themes in this book, viruses continue to provide clues to many biological processes that were not known only a few years ago—the phenomena whose significance eclipses host–pathogen interactions. Posttranscriptional gene silencing, RNA interference, and related phenomena—which were all but unknown 10 years ago and are, of course, all the rage of molecular biology now (Zamore and Haley, 2005) and have been recognized by 2006 Nobel Prize—were discovered and understood to be the RNA-level effects by plant virologists. The crucial observation was that RNA produced by virus-derived transgenes in plants is sufficient to shut down the infection by the same virus (Lindbo and Dougherty, 1992a,b, 2005). When the host genes
required for this shutdown were isolated, one of them turned out to be eukaryotic RNA-dependent RNA polymerase, which had been cloned originally as the enzyme responsible for viroid replication in tobacco (Schiebel et al., 1998). Most likely, this is the same activity that was discovered by Fraenkel-Conrat decades ago as he was trying to dissect the enzymology of virus replication in plants (Khan et al., 1986). There are multiple evolutionary and functional connections between replication of virus RNA and posttranscriptional gene silencing in eukaryotes (Ahlquist, 2002).

In the future, there will be more discoveries from the observations of viruses. The following is one example of what may be expected: Genomic RNA of brome mosaic virus, a plant virus that replicates in cytoplasm without a DNA intermediate, appears to contain modified bases, pseudouridylate and ribothymidylate (Baumstark and Ahlquist, 2001). We already knew that rRNAs and tRNAs are full of these and other modifications, but to find them in virus mRNA, even though the modified region is between two cistrons and is not translated, is a surprise. In the same vein, several plant viruses encode a domain homologous to the 2-oxoglutarate-dependent dioxygenase/AlkB family, which was discovered by comparative sequence analysis and is predicted to possess demethylase or dealkylase activity (Aravind and Koonin, 2001). Some of the members of this family appear to demethylate RNA (Ougland et al., 2004), although the activity of virus-encoded homologs has not been determined. I believe we are looking at the two facets of the enzymatic modification of mRNA: one equivalent to mutational damage, which needs to be repaired, and the other having a functional role, perhaps in control of mRNA stability, folding, and translation. If, as I expect, we will learn in the near future that cellular mRNAs also contain functionally important enzymatic modifications, perhaps constituting another level of regulation and recoding, the first indications will have come from viruses. (In a sense, discovery of the cap structure on eukaryotic mRNA is part of the same phenomenon, and this observation also came from studying viral transcripts; Wei and Moss, 1974; Wei et al., 1975; Ensinger et al., 1975; Keith and Fraenkel-Conrat, 1975).

But let us return to what this chapter is about—the role of viruses (and virologists) in defining comparative genomics. The reasons why viruses have been so popular as model systems are relatively simple: The number of genes in virus genomes is small, and the amount of genetically homogeneous progeny that can be obtained in the laboratory is large. These were the properties that made it possible to clone and sequence virus DNAs and RNAs early on. Even before the first genome sequence of a cellular life-form, Haemophilus influenzae, was determined in 1995, there were approximately 200 complete virus genomes in the databases (the exact number depends on how similar strains and isolates of the same virus are counted). By that time, virologists had already realized that virus genomes should be studied in a systematic way.

In 1971, David Baltimore of the Massachusetts Institute of Technology, currently at Caltech, provided the first system of viruses that was based on the diversity of forms of virus genomes and the modes of their expression (Baltimore, 1971). Whereas genomes of cellular organisms are, in a way, all the same—made of double-stranded DNAs, with all other nucleic acid forms being transient in the life cycle—viruses are all different. Not only do virions of different viruses contain all sorts of genetic material—DNA or RNA, single-stranded or double stranded—but also some virus life cycles do not include double-stranded DNA stage at all. This diversity needed to be rationalized, and Baltimore's system did just that. The fundamental idea was that every virus needs to produce a minimal set of two genetically encoded products ("sense-carrying units"): copies of the genomic nucleic acid and at least one mRNA to express proteins. The pathways sufficient to perform both tasks are schematically represented in Fig. 4.1.

Baltimore noted that "[t]here are many viruses about which so little is known that they cannot be placed in the scheme.... However, as their transcription becomes understood, either
they should fall into place in a recognized pattern, or new classes will have to be added."
What is most amazing is how little of such an addition was in fact required since 1971. In fact, about
the only finding that makes it to the classification now, 35 years later, is that some viruses with
single-stranded RNA genomes have ambisense segments (i.e., RNAs in which one part is
positive-sense and the other is negative-sense). This is a relatively minor adjustment; most
groups of DNA viruses also transcribe different portions of genome into separate, less-than-
genome-length mRNAs, which Baltimore did not fail to discuss in 1971.

Let us now consider Baltimore’s proposal from a broader point of view. Since ancient times,
people have been pursuing the quest for the natural system of living forms. We would like to
know whether comparison of genomes brings us closer to such a system.

We are looking for ways to rationally organize a large number of diverse objects. As a
rule, the objects have complex structure that is not completely known. A sensible start
would be to find some number of categories, such that every object could belong to one cat-
egory. It is good if the number of categories is much smaller that the number of objects. The
main way to assign the objects to categories is to find some sort of similarity between these
objects and to use the similarities to define groups of related objects, be it genes, viruses, or
anything else.

There are different ways to organize things by similarity. To proceed, let us define the
meaning of the following words: classification, systematics, taxonomy, and phylogeny. Instead
of reviewing the history of (changing) usage of each of the four words, I intend to stay as close
to their literal meanings as possible.

Classification should be about classes. Classes, or categories, of objects can be defined in any
way we wish—for example, by the shape of the objects, the second letter of their names, or any
other properties. Classes do not necessarily uncover any intrinsic properties of the objects that
we classify. With the appropriately chosen basis for our classification, however, classes may
turn out to be natural categories, representing some objective and essential properties of the
entities. Drawing a line between objective and arbitrary classifications is not always easy. For
example, grouping words by the first letter or grouping people by the last letter of their last
names seem quite arbitrary. Yet, most Russian words that start with “a” are borrowed (often
from Greek but sometimes from Arabic by way of other languages), whereas some people may
recognize my last name as Armenian. Therefore, at least in a small way, certain patterns reflect-
ing something important about words in a language (or, in this case, about historical trajecto-
ries of certain words) can be gleaned merely from grouping them into classes based on a simple
alphabetical rule.

Classification is more interesting if it is hierarchical, i.e., some categories are themselves
grouped into a category. Again, there is no constraint on the ways in which we build such
hierarchies.

Figure 4.1. David Baltimore’s system of viruses. Reproduced from Baltimore (1971) by permission of the
American Society for Microbiology.
Systematics should be about systems. This means not only that we group objects into classes but also that we would like these classes to have relationships to each other, which represent some intrinsic properties of these classes. The order of letters in the alphabet does not seem to reveal anything deep and intrinsic about them; the alphabetical classification, therefore, is not a systematics because there is no specific relationship between words starting with “a” and words starting with “b.” We want something better—some guiding principle on which to build the system.

Biologists would say, however, “surely such a guiding principle has been discovered; it is called biological evolution.” Indeed, evolution is the force that produced and shaped the observed life, and most approaches to biological systematics essentially amount to uncovering the evolutionary history of the taxa that are examined. But there is no reason why the historical development of organisms should be the only guiding principle for systematics. An analogy from chemistry is the Periodic System of Elements. Conceived in 1869 by Dmitry Ivanovich Mendeleev, the system captures the majority of then-known physical and chemical properties of elements and their similarities and consistent differences between different groups of them (Mendelejeff, 1869). Importantly, many properties of each element are determined by, or at least strongly correlated with, a single parameter. This parameter later turned out to be the positive charge of the nucleus of a chemical element, which, in Mendeleev’s times, could be only approximated by atomic weight. Interestingly, the contemporaries of Mendeleev, notably Newlands in England and Meyer in Germany, developed similar systems at the same time. However, a notable distinction of Mendeleev’s system is that, unlike the proposals of Newlands or Meyer, it contained empty classes (i.e., places for the elements that remained to be discovered). Therefore, the system is robustly organized by a guiding principle that reflects something profound about the elements and even predicts new things about chemical organization of the universe. It is systematics, under the literal definition proposed here. Yet, it does not tell us much about the natural history of chemical elements.

Biological species are different from chemical elements, of course. The main distinction of biological species from most other things is the foundation of the sense-carrying units, evolving by descent with divergence from the common ancestor. Even so, there is no reason why some important aspects of form and diversity of living things should not be presented as some sort of a “periodic system.” Baltimore’s system of viruses is a good example of just such a representation.

Taxonomy should be about taxa. That is where evolution starts playing a more prominent role. Taxa are groups nested in an hierarchy. In biological systems, the most common and natural reason for the existence of a hierarchy is evolutionary process. For some thinkers, all taxonomy is by definition evolutionary. On the other hand, biologists often make use of nonevolutionary hierarchies. One example is EC, the enzyme classification employed by the International Union of Pure and Applied Chemistry (Tipton and Boyce, 2000). EC divides all enzymes into six groups—oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases—each of which is hierarchically divided further. For example, EC 2, transferases, includes group EC 2.7, enzymes that transfer phosphorus-containing group. EC 2.7 includes EC 2.7.4, phosphotransferases with a phosphate group as acceptor, which has a member EC 2.7.4.2, phosphomevalonate kinase. There exist at least two enzymes with phosphomevalonate kinase activity that have different, unrelated sequences and probably do not share any common ancestor (such isofunctional but unrelated enzymes are discussed in Chapter 5). Thus, EC may (although some will say should not) be called a biological taxonomy, but it is not an evolutionary taxonomy.

More recently, the genomics community started putting together Gene Ontology, a knowledge base and controlled vocabulary for annotating gene function. This is also a hierarchy that contains both evolutionary and substantial nonevolutionary components.
Finally, phylogeny should be about genesis of phylae. Although phylae literally means "races" or "classes" in Greek, the scientific meaning of the word has to do with branches in the trees that depict historic relationships between species. Although phylogenetic trees are familiar to a biologist, trees are also objects of mathematics. They are formally defined in graphs, and there are many different ways to construct tree-like graphs. Not every tree is truly representative of the evolutionary history of the species that we are studying (see Chapter 11). Moreover, even if the objects of interest do not have any evolutionary relationship, we can still construct a tree-like representation of connections between them; this is common, for example, in comparing gene expression patterns and other genomewide numeric data (see Chapter 14). Thus, phylogeny is a tree-like representation of ancestral relationships, but not every tree-like representation of biologically interesting information is a phylogeny.

With this, as literal as possible, understanding of classification, systematics, taxonomy, and phylogeny in hand, let us examine Baltimore's proposal once again. The scheme shown in Fig. 4.1 surely is a classification, and it is also a system: Not only are the objects (viruses) partitioned into classes but also this is done on the basis of a principle. As with any good system, the property chosen as the basis of the system—in this case, the path from genomic nucleic acid to mRNA—allows one to make many predictions of other properties of viruses. One such prediction, discussed by Baltimore, was that virions that do not pack mRNA have to rely on cellular machinery to produce it, or to pack virus-encoded transcription enzyme into virions. Baltimore's scheme, however, is not a taxonomy—there are no nested taxa in it. It is also not a phylogeny. The word "evolution" is not mentioned in the paper at all; even its root ("evolved") is only used once, in passing. I do not suppose this indicates lack of interest in evolution on Baltimore's part; more likely, he did not believe there was enough evidence to suggest a sensible scenario for the evolutionary origin of different virus groups. Indeed, as we will soon see, the evolutionary relationships between Baltimore's groups are not intuitive.

Baltimore also remarked, "Viruses with similar transcriptional systems could have different replicational systems, leading to the necessity to extend the class designations." Such an extended system proposed a few years later by Vadim Agol (1974) of Moscow University and the Institute of Poliomielitis was the next major step in comparative genomics.

Agol defined four types of genetic elements: (+)DNA, (-)DNA, (+)RNA and (-)RNA. Eight "elementary acts of synthesis" are theoretically possible, two for each of these genetics elements; for example, (+)DNA can be copied either into (-)DNA or (-)RNA. There are 44 distinct "full acts of synthesis," or interconversions of single-stranded and double-stranded molecules. If we require that each life cycle contains one act of mRNA synthesis and at least one act of multiplicative synthesis (increase in genome copy number), and only consider graphs with no more than three edges, there are 35 distinct life cycle graphs that can be constructed from these elements (Fig. 4.2).

The constraint on the number of edges was introduced for convenience; no fundamental reason is known as to why the life cycles of viruses should have only two or three edges. In fact, viruses appear to follow this rule: By and large, the life cycles of all known viruses are covered by this set of simple graphs. (A few cases in which four or five edges may give a better explanation of virus reproduction mechanism are discussed in Agol's work, but with reasonable, minor simplifications, they are all reduced to three-edged graphs.)

Agol's scheme highlighted some empty classes and predicted that virus life cycles corresponding to some of these new classes will be discovered. In fact, the class A1 in the DDRD type has since been filled. This class is represented by animal hepadnaviruses and two groups of plant pararetroviruses—caulimoviruses and badnaviruses. Interestingly, this class was identified by the author as one of those for which such future discovery was most likely, based on the observation that all acts of synthesis required for this class were already discovered in nature.
Figure 4.2. Vadim Agol's system of viruses. Polarity of DNA (D) and RNA (R) strands is shown next to each nucleic acid. The nucleic acid that is part of a virion is on the left side of each circular graph. If the cycle does not contain the (+)RNA stage, an mRNA needs to be synthesized in an additional step. Such cases are shown by the linear arrows pointing outside the circles. Shaded graphs are the strategies that had been discovered before Agol's work, and the lightly shaded graph at the bottom left is the life cycle predicted by Agol to exist and indeed discovered later. Modified from BioSystems, 6, Agol, V. I., Towards the system of viruses, pp. 113–132, Copyright 1974, with permission from Elsevier.

Agol's scheme is a classification and also a system. It is a taxonomy as well, because it includes a natural hierarchy of classes, superclasses, and types. On the other hand, this hierarchy is not phylogeny: In recent years, it became clear that some of the types, particular RR and DDRD, are evolutionarily related, and that the DDR type may have to be split into evolutionarily independent lineages. Baltimore's classes are distributed among three of Agol's types. For example, Baltimore's classes III, IV, and V belong to Agol's type RR, and Baltimore's class IV is one of the six classes in type DDRD. This is because Baltimore was mostly interested in classifying the existing viruses, whereas Agol was concerned in an exhaustive enumeration of all logical possibilities, and his theoretical system was therefore set up to include "blank" classes. Many graphs remain purely theoretical possibilities 30 years later. The aforementioned class DDRD-A1 is the only novel class discovered since then; three of the highest level groups, types DRRD, DRD, and RRD, remain completely vacant, and approximately half of the classes in other types are also empty.
The large number of “still-empty” cells in Agol’s system is interesting because it indicates that classes of biological objects are usually unequally populated. In other words, nothing in biology is purely combinatorial: Some sets of properties characterize a very large number of biological objects, other combinations of properties are found rarely, and there are many seemingly plausible but nonetheless empty classes. It is almost too easy to discover the basic components of living systems and to invent the ways of mixing and matching them. What is much more difficult is to understand the constraints that are imposed on such combinations in the evolution of life. Most of the time, we can only guess about the reasons that “forbid” some of the combinations. In fact, in the case of Agol’s empty types, his explanation was very good: The empty superclasses typically involve an mRNA synthesis directed either by a single-stranded DNA or by an RNA–DNA hybrid, and all empty types involve a synthesis of a (+)DNA strand on a (−)RNA strand, or the opposite (+)RNA → (−)DNA reaction. In the environment of the double-stranded DNA genomes, the RNA strands of these duplexes will be prone to destruction, and single-stranded DNAs will be restored to two strands by the DNA repair system. This remains a guess; experimental testing of this suggestion will require the construction of artificial viruses, which appears to be within reach (Cello et al., 2002; Smith et al., 2003).

The systems of viruses described previously mark the beginning of comparative genomics in a most direct sense—that is, the work of comparing different genomes. Thus came to be the idea of virus genomes as complete multigene entities, which are related to each other in specific ways and can be studied as a whole. In the late 1970s and early 1980s, sequences of individual virus genes and of complete virus genomes started to accumulate, and at approximately the same time, biologists started to get better access to computing.

In 1980, David Botstein of MIT (now at Princeton University) presented a metaphor that continues to catch on in molecular evolution and comparative genomics, in a work called “A Theory of Modular Evolution for Bacteriophages.” That work was based on the observations of several temperate bacteriophages with double-stranded DNA genomes. Only partial nucleotide sequences were known for these phages, but genetic maps of many of them, particularly the lambda phage and its close relatives, were worked out in great detail. A remarkable feature of these genomes was that the genes involved in one and the same function were, more often than not, positioned close to each other in phage genomes, and arrays of such genes occupied similar positions within the genomes. For example, in several lambdoid phages, the group of genes coding for the phage head component was followed by the group of tail genes. After the tail genes, a stretch of DNA could be found to which no gene functions were mapped at the time, but remarkably, this stretch accounted for roughly the same fraction of each genome. This was followed by genes involved in DNA recombination, then by DNA replication factors, and, finally, by lysis genes (Botstein, 1980).

Why was the conserved gene order remarkable? Colinearity of many genes in cellular life-forms had already been demonstrated by genetic mapping of related enterobacteria, such as Escherichia and Salmonella. In phages, however, there was a crucial difference: In some cases, only the locations of the isofunctional genes were conserved, whereas the gene sequences were quite different. One example was phage tails that look completely unlike one another. Lambda’s tail is long and flexible and the tail of P22 is short and rigid, yet both are encoded by tail modules located at identical positions in these phage genomes. Another example is lack of similarity between biochemical activities of lysis genes in phages 80 and P22, respectively represented by glycosidase and endopeptidase, which target different chemical bonds within peptidoglycan of the cell wall. In order to explain this contrast between gene colinearity and lack of molecular similarity between functional groups of genes, Botstein proposed that these groups can be transferred between genomes independently of one another, which was in agreement with the observation that different lambdoid phages could recombine in multiple spots.
along the genome. This ability, then, provided the mechanism for exchanging genes or, more important, groups of genes or "modules"—for example, the complete set of a half dozen genes required for tail assembly.

Thus, the "module" idea has been released into comparative virology, from where it passed to other areas of biology and, in due course, to the emerging discipline of comparative genomics. It is worth remembering, however, that as early as the late 1950s, the idea of bacterial operons (i.e., groups of functionally linked genes in bacteria that are also adjacent on a chromosome, ensuring ease of regulation by way of a single control element) had been proposed (Jacob et al., 1960). Dozens of operons in bacteria and in bacteriophages were known by 1980. In fact, the choice between lysis and lysogeny pathways in temperate bacteriophages was one of the favorite models to study operon organization of genes and transcription regulation in prokaryotic cells. So, what was the difference between operons and Botstein's modules?

One distinction is that operon theory focused mostly on the functional aspects of gene clustering on a chromosome—adjoining genes in bacteria are easier to coregulate, using elaborations of the basic scheme that involves a single activator of transcription (operator) and expressing multiple genes by internal initiation of translation on polycistronic mRNA. The theory of modules considers the same gene clusters and emphasizes the ease with which they can be transferred between genomes by DNA recombination. Although evolutionary mobility of operons had been discussed before Botstein's work, it had been viewed as a relatively rare event, occurring in the background of much larger bacterial genome, which was believed to evolve mostly by mutational divergence and by recombination that involves long regions of high sequence similarity. In contrast, Botstein's proposal decomposed the entire phage genome into a moderate number of parts that could be inherited independently of one another—there was almost no "stable background." Botstein (1980) stated, "A rather large and apparently diverse group of temperate bacteriophages are related in ways not easily accounted for by standard ideas of evolution along branching trees of linear descents." Thus, evolution of viruses may have to be depicted as something different than a tree. Notably, this idea derived from the empirical observation of molecular-genetic characters in completely mapped, if not yet sequenced, genomes.

In Botstein's proposal, the relatively short regions of high nucleotide sequence similarity, located at the junctions between different modules, were thought to be necessary for the mechanism of module exchange. Nowadays, many DNA recombination pathways are known that do not require extended complementarity: When DNA homology is present, the recombination machinery will take advantage of it, but when there is no lengthy stretches of identical DNA, recombination may occur anyway. Recombination between distantly related DNA genomes may have played an exceptional role in evolution of life, especially early on (see Chapter 11). Genomes of the RNA viruses also contain plenty of evidence of gene exchange.

Modular evolution theory fits well with Pauling and Zuckerkandl's idea that a molecular function can be performed in several mechanistically and evolutionarily unrelated ways. Functionally analogous proteins do indeed exist, as the example with lysis genes of temperate phages shows. Functional convergence at the molecular level will be discussed in much more detail in Chapter 6.

Finally, the concept of modular evolution indicates that morphology may not always be a reliable guide in evolutionary studies. The type of phage tail has been often used as a phylogenetic marker, but if tail genes constitute just one module among many, accounting for only a fraction of all phage genes, and if this module is free to "mix and match" with other modules, then two phages with similar sets of tail genes may be placed into the same taxonomic
group regardless of the other genes they have. On the other hand, phages that share most of the genome but have different sets of tail genes may fall into separate groups.

One could argue that the situation is perhaps different when we deal with morphology of cellular (especially multicellular) organisms. Here, the hope may be that morphology is determined by interaction of many genes and reflects their joint presence and coevolution—in contrast to viruses, in which morphology may be determined by too few genes and be less representative of the rest of the genome. The International Committee on Taxonomy of Viruses, nonetheless, still relies heavily on its virion morphology for defining at least the higher order virus groups. This may be satisfactory for the purposes of classification and taxonomy, but it would hardly be enough if we want our taxonomy to be informed by evolution and to represent phylogenetic relationships. The impact of genome modularity on phylogeny is discussed in Chapters 11 and 12.

These are just some of the implications of the “modular theory”: Chapter 14 provides a brief discussion of “modular biology,” a new proposal for organizing and studying genomewide data. But, perhaps surprisingly, the beginnings of this project are in sequencing virus genomes.

Another nonphylogenetic approach to virus classification was put forward by Eugene Koonin in 1991 when he was still at the Institute of Microbiology, Academy of Sciences of the USSR (Koonin, 1991). That work, dealing only with positive-stranded RNA viruses, had already benefited from sequencing of virus genomes, but it examined not so much the individual sequence relationships but, rather, similarities in genome strategies (i.e., the molecular mechanisms employed by viruses to express their genes and replicate their genomes). The topic was thoroughly familiar to Baltimore in 1971 and Agol in 1974, but they were mostly interested in the strategies of viral replication and transcription. However, for the largest group of viruses, those of the RR type, the diversification of molecular strategies of genome expression takes place at the level of mRNA translation into proteins.

The main outcome of virus genome expression is the production of individual virus proteins. All viruses usurp ribosomal machinery of the host cell in order to express their proteins, and almost all viruses encode more than one mature protein. Therefore, the main task of a virus expression strategy is to produce several protein species from their mRNAs.

In bacteria, all virus genes can be translated into separate proteins from one polycistronic RNA, which is the same as genomic RNA (the RNA that is translated and the RNA that is packaged into virions may not be physically the same molecule, but, as a first approximation, they have the same nucleotide sequence). Translation of many open reading frames (ORFs) from the same polycistronic mRNA is the first expression strategy.

Many viruses use another mechanism to produce individual virus proteins: They code for a large precursor protein that can be proteolytically processed into fragments with distinct roles in the virus life cycle. Usually, the proteases required for such processing are included in the large protein precursor and are able to release themselves as well as process the rest of the protein (Bazan and Fletterick, 1989; Gorbalenya et al., 1988, 1989, 1991).

The third strategy of expressing individual proteins is to produce several different mRNAs by transcribing the minus strand of the polycistronic RNA genome. The full-length mRNAs will serve as the genomic RNA that can be encapsidated in the progeny virions and possibly also as the messenger for translation of the 5’-proximal ORF. But (−)RNA may also be transcribed, starting at some internal position, into a (+)RNA that has the same 3’ terminus as the genomic RNA but is less than genome length. The role of such subgenomic RNAs is to direct translation of the downstream genes in the virus genome.

Finally, and quite trivially, a virus can possess a fragmented genome, in which each RNA fragment encodes exactly one protein that does not need to be processed further. All possible
combinations of the four mechanisms (polycistronic expression, processing of a polyprotein precursor, synthesis of subgenomic RNAs, and genome fragmentation) give 15 possible strategies (Table 4.1).

Koonin also examined, but decided not to include into his classification, such translation-level events as frameshift, readthrough of leaky termination codon, and other recoding mechanisms. This "ribosome gymnastics" is popular among positive-stranded RNA viruses and also among distantly related retroviruses and some viruses with double-stranded RNA genomes. Another discovery in recent years is the discontinuous synthesis of 3'-coterminal RNAs in some virus groups (e.g., coronaviruses); it may be sufficiently different from other mechanisms of subgenomic RNA formation and could be placed into a separate column.

In truth, only 9, rather than 15, expression strategies were proposed in Koonin's 1991 paper, and accordingly there were 18 classes. All those classes that involve polycistronic translation were collapsed into just 2, with and without VPg. The reason was that for a long time, most people thought that polycistronic translation was unavailable in eukaryotic cells. There,

| Genome expression mechanisms | Genome segmentation | NonVPg utilizing replication | VPg-utilizing replication |
|-----------------------------|---------------------|------------------------------|---------------------------|
| 1 Polycistronic expression (1) | Segmented           | ssRNA phages                 |                           |
| 2 Polyprotein processing (2) | Nonsegmented        |                             |                           |
| 3 Subgenomic RNA formation (3) | Segmented           | Flaviviruses                 |                           |
|                             |                     | Pestiviruses                 |                           |
|                             |                     | Dianthoviruses               |                           |
|                             |                     | Tobraviruses                 |                           |
|                             |                     | Tricornaviruses              |                           |
|                             |                     | Hordeiviruses                |                           |
|                             |                     | Carmoviruses                 |                           |
|                             |                     | Tombusviruses                |                           |
|                             |                     | Potexviruses                 |                           |
|                             |                     | Carlaviruses                 |                           |
|                             |                     | (Capilloviruses)             |                           |
| 7 (1)+(2)                   | Segmented           | Engineered flavivirus        |                           |
|                             |                     | derivatives                 |                           |
| 8                           | Nonsegmented        |                             |                           |
| 9 (1)+(3)                   | Segmented           | Nodaviruses                  |                           |
| 10                          | Nonsegmented        | Alphaviruses                 |                           |
| 11 (2)+(3)                  | Segmented           | Rubiviruses                  |                           |
| 12                          | Nonsegmented        | Coronaviruses                |                           |
| 13 (1)+(2)+(3)              | Segmented           | Tymoviruses                  |                           |
| 14                          | Nonsegmented        |                             |                           |
| 15                          | Genome segmentation |                             |                           |

Table 4.1. Classification of Genome Replication and Expression Strategies in Viruses with Positive-Stranded RNA Genome

*Modified from Koonin, E. (1991). Genome replication/expression strategies of positive strand RNA viruses: A simple version of a combinatorial classification and prediction of new strategies. *Virus Genes* 5, 273–282. With kind permission of Springer Science and Business Media.

*Engineered construct corresponding to one predicted class of viruses.
translation initiation most often proceeds by ribosome scanning from the 5' end, which provides for efficient translation only of the ORF closest to the 5' end of mRNA. Or so it was believed for several decades, until evidence of many exceptions to this rule was discovered; most notably, several viral and some cellular mRNAs have sequence segments to which ribosome can bind directly, without scanning from the 5' end, and to initiate translation on an ORF placed downstream of such element (called IRES). In fairness to the scanning mechanism, it has to be noted that these control elements seem to be used in nature mostly for scanning-independent expression of the 5'-terminal ORF and not for polycistronic expression. However, artificial bicistronic and polycistronic eukaryotic messengers have been constructed, in which each ORF is placed under the control of its own IRES element and all ORFs in such constructs can be expressed. Therefore, it seems possible that sooner or later we will find a eukaryotic virus that also uses this expression mechanism.

The mechanisms of RNA replication are less well studied (although significant progress has been made in this field since 1991; see Ahlquist et al., 2005). Koonin sensibly chose the mechanism of initiation of RNA synthesis as a character that can have two states, namely use of a protein (called VPg) with a covalently attached nucleotide as the primer of RNA synthesis or VPg-free initiation. Combination of these two modes with 15 modes of individual protein generation gives 30 strategies (Table 4.1).

Fifteen years after Koonin's study, the state of the affairs has not changed much. Despite ongoing sequencing of virus genomes and the discovery of several novel virus taxa, empty cells mostly remain empty: Replication/expression strategies of newly described virus groups tend to fall into already occupied classes. However, the principal possibility for the existence of some currently empty classes has been demonstrated by genetic engineering (Geigenmuller-Gnirke et al., 1991). But correlations observed by Koonin in 1991 mostly retain their status: Prokaryotic-style polycistronic expression is actualized only by RNA phages and remains a theoretical possibility for eukaryotic viruses; RNA phages remain the smallest and most homogeneous group, not known to rely on VPg-primed RNA replication; and, for reasons that are unclear, VPg-dependent replication, which is common among eukaryotic viruses, is closely associated with the polyprotein processing strategy.

In summary, Koonin's combinatorial scheme is a classification and a system but not really a taxonomy. As indicated by Koonin, it also does not capture phylogenetic signal very well. Although the same class often contains viruses that are evolutionarily closely related, the opposite is not true: Some groups of viruses that are closely related at the sequence level may have quite different strategies of replication/expression.

A more general conclusion is again the same as before: The combinatorial approach to classification (at least to virus classification) appears to produce a larger number of possibilities than is actually employed by nature. This results in many empty classes in Agol's scheme and in Koonin's classification. Apparently, the evolutionary process operates under constraints, so its results do not look like the product of indiscriminate mixing and matching. Perhaps every combinatorial classification should be expected to contain many empty classes (more examples of this will be provided when the patterns of the presence and absence of genes in genomes are discussed in Chapters 6 and 11).

It is now time to examine another line of comparative virus genomics, namely the study of evolution of individual protein sequences. In the 1980s, these studies began producing important, if not quite expected, results.

Three biochemical activities play a major role in genome replication and expression of positive-stranded RNA viruses (Baltimore's class IV; note that some of the comparisons that we are about to discuss will show that Baltimore's classes III and VI share a common ancestor with class IV). First, RNA strand synthesis on an RNA template requires the processive nucleotidyltransferase activity provided by the enzyme RNA-dependent RNA polymerase.
Second, many positive-stranded RNA genomes encode another enzyme, RNA-dependent ATPase, which plays a nucleic acid remodeling role (e.g., removal of secondary structure from self-paired regions of RNA or, conversely, RNA duplex formation), and may also control association and dissociation of proteins with RNA. Viral (and homologous cellular) proteins capable of doing all this are known under a slightly imprecise name—helicases. Third, cleavage of virus polyproteins requires virus-specific proteases, which belong to several unrelated classes. In addition to these enzymes, viruses code for capsid proteins. In the case of positive-stranded RNA viruses, there is often just one such protein, or there may be several, either expressed independently or produced by proteolysis of a larger precursor, sometimes while or after the capsid shell is assembled. (Other classes of virus-specific proteins—some possessing interesting enzymatic activities and others playing regulatory roles in virus life cycle and host evasion—are less widespread and are not examined here).

The relationships between various virus proteins, and between them and their cellular homologs, began to come to light in the early 1980s. Everything started falling into place in 1984. One the most important activities, template-dependent RNA synthesis, has been genetically mapped to a specific translation product, or at least to a specific RNA segment, in genomes of several viruses. Kamer, Argos, and co-workers (Kamer and Argos, 1984; Argos et al., 1984) reported statistically significant sequence similarity between what was (correctly) inferred to be the main virus RNA replication enzyme RNA-dependent RNA polymerase (RdRp). At approximately the same time, Haseloff and co-workers (1984) compared a subset of these proteins and noticed the same similarities but also additional relationships in two other protein domains.

The region of the highest sequence conservation in RdRp enzymes contained a string of approximately 10 residues, with the tripeptide GDD in the middle (Fig. 4.3), often preceded by Y or another bulky hydrophobic residue. Some positive-stranded RNA viruses (e.g., coronaviruses sequenced later) have SDD instead of GDD. RNA viruses from other groups, those with double-stranded and negative-stranded RNA genomes as well as those that employ reverse transcription, all possess related enzymes, now known to perform processive synthesis of virus nucleic acid. The most conserved site in these proteins is always a variation on the “GDD” theme, typically taking the form of YxDD in retrovirus viruses and UUDD (where U stands for a bulky hydrophobic residue) in negative-stranded RNA viruses. The active center of RdRp contains two magnesium ions that play a direct role in catalysis, and the first of the two aspartic acid residue as well as some of the preceding residues (tyrosine in the best studied reverse transcriptases) play a direct role in interaction with these metal cofactors.

Common features of these enzymes are not limited to the 10-residue region: There are several other areas with especially high similarity, including some positions where the residues were the same in all sequences. In addition, the approximate distances between these conserved sites were similar in all sequences, and the positions of the RdRp-related regions as a whole within each virus genome were partially conserved. This alignment is quite different from what was considered, at the time, to be typical sequence conservation in a protein family.

The existence of families that consist of homologous proteins (orthologs in different species or paralogs in the same species; see Chapter 3) was well established during the 1960s and 1970s. However, the methods of finding homologs were less sensitive than what we have today, and they tended to recover proteins that were globally similar to one another. Despite many point mutations, and occasional insertions/deletions, each sequence included in these alignments mostly consists of segments that were clearly aligned to their counterparts in all other sequences, and the percentage of residues that fall into conserved regions was quite high.

This does not mean, of course, that the existence of more remote relationships between proteins was completely unsuspected. Not only did theoretical consideration of descent with divergence suggest that such relationships would exist but there was also plenty of empirical
Getting Ready for the Era of Comparative Genomics

Figure 4.3. Conserved motifs in virus RNA-dependent RNA polymerases. Reprinted from Kamer, G., and Argos, P. (1984). Primary structural comparison of RNA-dependent polymerases from plant, animal and bacterial viruses. Nucleic Acids Res. 12, 7269–7282, by permission of Oxford University Press.

evidence—morphological, cytological, and biochemical—for the common ancestors of various taxa and, most likely, of all living things. In more doubt, however, was the sensitivity of our methods and whether we would ever learn to distinguish extreme evolutionary divergence from convergence or from random coincidence.

In the meantime, however, unexpected sequence relationships were observed increasingly more often. In the early 1980s, genomes of cancer-causing retroviruses were sequenced, and when sequences of their oncogenes were compared to sequence databases, they turned out to have homologs with known functions encoded by cellular genomes (Barker and Dayhoff, 1982; Doolittle et al., 1983). However, even though these similarities were unexpected, they were not remote but, in fact, quite high—so high that no statistical theory was needed to convince the audience that they were real indications of evolutionary relationship and similar molecular function of viral oncogenes and their cellular homologs. Thus, we were already doing well with “unexpected but high” sequence similarities; the problem was validating similarities that were plausibly expected and yet quite low.

The auxiliary evidence, however, strongly supported sequence similarities between proteins conserved in RNA viruses. First, there was plenty of genetic data mapping replication ability to the RdRp domain. Second, conservation of gene order in virus genomes, pointed out by Botstein in 1980, was likewise observed in RNA viruses. By gene order, we understand both the colinearity of genetically mapped functions and the conservation of several different domain
sequences, even if each such pair of domains is only moderately conserved. Third, the rapid growth of sequence databases became an important factor: One day we could be looking at two, possibly homologous, sequences, trying and failing to distinguish short regions of similarity signal from the background noise, and the next day we would see a new entry in the database that produces statistically significant matches to both of the sequences. This nontransitivity of sequence searching and scoring, and some ways to overcome it, was discussed in Chapter 2.

Technical details of all these comparisons are not very important to us at the moment. Nowadays, we have much more sensitive and rigorous methods of database search and statistical evaluation of similarities than those available 20 years ago, and homology between many virus RdRp enzymes can be established in an unbiased fashion. But remarkably, even today, some of the distant similarities between members of the RdRp family and their distant relatives are still difficult to detect by a casual BLAST search, and more involved, iterative probabilistic searches are required to match them. All the more glory to the pioneers of comparative molecular virology, who were able to unravel these similarities using the imperfect tools.

Which brings me, again, to the title of this chapter, claiming that something about the analysis of viral proteins turned out to be important not only for virology but also for comparative genomics in general. What was it? First, there is the aforementioned notion that gene orthology is not synonymous with high sequence conservation—some pairs of orthologs are closely related, whereas others are more distant—and analysis of virus-encoded enzymes and structural proteins was one of the first case studies that has drawn our attention to this. Second, not only is the identity between virus homologs low on average but also, as the proteins evolve, the similarity between orthologs becomes confined to the increasingly shorter fraction of the protein length—that turned out later to represent the mainstream way in which protein families evolve. Third, viruses were the first model of a very long evolutionary process. This obviously, does not have to be long in absolute time—more important is the rate of change per generation, and in the case of viruses, generations are short and the number of changes per genome per generation is higher than in cellular genomes—again, with obvious parallels to subsequent analysis of cellular DNA genomes.

Rapid evolution of viruses with RNA genomes is sometimes attributed to a high rate of nucleotide misincorporation in reactions catalyzed by RNA-dependent polymerases (although, of course, high mutation rate will not automatically translate into high evolution rate; see Koonin and Gorbatenya, 1989). But despite this handy molecular explanation, the pattern of conservation in virus-encoded proteins could not be easily dismissed as yet another unusual, extreme adaptation to intercellular parasitism. On the contrary, further sequencing of viral and nonviral genomes alike provided more of the same, and a portrait of a “typical” protein family started to emerge, looking like an assembly of sequence motifs separated by noisy linker regions. A previously more familiar pattern of high sequence conservation along the full length of the aligned proteins may in fact be a special, extreme case of sequence similarity, whereas the “virus-like” conservation, centered on most important sequence motifs, became a new null hypothesis of evolution within the protein family.

At approximately the same time as the observations of the unity of distantly related RdRp enzymes, two other protein families encoded by RNA viruses came to light. One such family consisted of cysteine proteases encoded by genomes of animal picornaviruses and plant comoviruses (Argos et al., 1984). The other family included RNA helicase, at that time known as a putative protein that was commonly found next to the RdRp domains and most likely was involved in replication. Both families displayed the same picture of significant sequence divergence as RdRp, with several relatively short conserved regions interspersed with longer regions where similarity could not be easily detected. Importantly, within a few years, each of these two families was connected to a particular family of cellular enzymes, which themselves showed extreme sequence divergence.
The relationship between cellular helicases and the second most conserved protein of RNA viruses was understood not all at once. In fact, the report of a large ATPase superfamily, of which helicases are a part, was first published in 1982 (Walker et al., 1982). In 1988, Hodgman published the alignment of helicases in the form of a series of conserved blocks along the sequence, removing from consideration the portions of proteins that were not amenable to proper alignment (Fig. 4.4; this was one of the first printings of sequence conservation in such condensed format, which remains popular today). So much for the idea that low conservation and short motifs are the property of virus proteins only.

The other family also extended beyond viruses. The title of an article, “Poliovirus-Encoded Proteinase 3C: A Possible Evolutionary Link between Cellular Serine and Cysteine Proteinase Families” (Gorbalenya et al., 1986), speaks to the significance of sequence comparisons of virus enzymes: Not only are viral and cellular enzymes distantly related but also their similarities illuminate the evolutionary relationships between different classes of cellular enzymes. In this case, again, the most pertinent information was presented as two short blocks of local sequence conservation.

These reports were followed throughout the next two decades by increasingly sensitive sequence analysis, resulting in many other conserved virus domains shared with prokaryotes and eukaryotes (Aravind and Koonin, 2001; Putics et al., 2005). In almost every case, the extent of sequence variation among virus homologs was comparable to variation in the members of the same family encoded by cellular organisms, confirming that the extreme divergence between viral proteins is not a fringe phenomenon.

| Motif | I | II | III | IV |
|-------|---|----|-----|----|
| uvzD  | 26 VLAGAGSGKTRVLV_174_NILVDEFQNTN_16_VMIVGDDDSIQY_26_QNYRSTSI |
| rop   | 19 VLAGAGSGKTRVT_175_YLLVDEVQDTN_16_FTVVGDSSSIY_26_QNYTSSGRI |
| recB  | 20 IASAGTGKTPITA_345_VAMIDEFQDTD_18_LLLIGDPKQAIY_24_TNNWSPAPGM |
| recD  | 164 ISGGPGGTKTTTV_82_VLVDASAMID_16_VIFLDDRDQAS_24_QLSRILGTH |
| EBV   | 69 ITGTAAGBSTTSSV_113_VIVVDEAGTLS_26_IVCVGSPQTQA_44_NNKRCTDVQ |
| HCMV  | 117 VTGTAAGBSTTSSV_113_VIVVDEAGTLS_26_IVCVGSPQTQA_44_NNKRCTDVQ |
| HSV   | 94 ITGNAAGSGKSTCVQ_136_VIVDEAGLLG_26_IVCVGSPQTAS_44_NNKRCEVHE |
| VZV   | 87 IGSNAGSKSCID_135_VIVDEAGLLG_26_IVCVGSPQTDS_44_NNKRQEDDD |
| PIF   | 255 YTGAGTGGKSILLR_46_ALVDEASMLD_25_LIFCGDFQQLPP_29_KVFRQRGDV |
| A1MV  | 821 VDGVACGCKTNIK_55_RLIDECFLQH_15_VIGFQDLTIIPF_22_ITWRSADA |
| BMV   | 687 VDGVACGCKTIAK_54_RLIDECFLQH_15_VIGFQDLTIIPF_22_ITWRSADA |
| CMV   | 709 VDGVGCCKTIAK_54_RLIDECFLQH_15_VIGFQDLTIIPF_22_ITWRSADA |
| TMV   | 829 VDGVGCCKTIEIL_57_RLIDECFLQH_15_VIGFQDLTIIPF_22_ITWRSADA |
| TRV   | 901 VDGVCCKSTMIIV_56_VLHFDEAŁMAH_15_CIČQIDONQISIF_24_ETVRSADV |
| SFV   | 183 VEGVPGSGKSAIIX_110_LVIDVEAFACH_16_VLVDGDPQCGF_21_ISRTRTPV |
| SV    | 183 VIGTPGSGKSAIIX_110_LVIDVEAFACH_16_VLVDGDPQCGF_21_ISRTRTPV |
| IBV   | 1209 VGQPGSGKSHFAY_50_LVIDVEFSLMT_15_VYVGDFQAQLPA_30_KCYRCPKE |
| BNYVV1| 893 VKGPGGTGKSPFLR_48_IFVDEFTAYD_11_YLVDGDEQQTGI_25_MNFRNPVHD |
| BNYVV2| 121 VLGAPCVKSTSIK_49_TMLVDEVTRVH_11_VICFGPDAQGQIL_19ARSERFGKAT |
| BSMV  | 267 ISVPGSGKSTIVR_41_LVIDVEFYTLAE_11_VLVDGVAQQKA_18_TYRLGQET |

Figure 4.4. Fragment of multiple sequence alignment showing the most conserved motifs in putative viral replication enzymes with helicase or DNA/RNA-dependent ATPase activity and in their cellular homologs. Modified from Hodgman (1988) by permission of Nature Publications, Inc.
For the most important viral enzyme, RNA-dependent RNA polymerase, the discovery of true cellular homologs has taken the longest time. It was known since the early 1980s that the most conserved regions of virus RNA-dependent polymerases contain the universally preserved diaspartic DD. Even though many other polymerases and nucleotideyltransferases also contain aspartic acid residues in their active centers, the sequence comparisons have unequivocally supported only the relationship between RdRp and RNA-dependent DNA polymerase, the replication enzyme of retroviruses and related retroelements. The latter, of course, are found in abundance in eukaryotic genomes and, to lesser extent, in prokaryotes. So the homologs of RdRp encoded by cellular genomes, or at least the prime suspects for this role, were known for a while. When x-ray structures of RdRp and reverse transcriptases became available, their close similarity and equivalent positions of the most conserved residues left little doubt that the two classes of polymerases are homologous. The shape of reverse transcriptase resembled the right hand, with “palm,” “fingers,” and “thumb” domains; the palm domain contained the residues required for nucleotideyltransferase activity. However, the cell-encoded reverse transcriptases were all associated with integrated proviruses, retrotransposons, and other such elements that seemed to be genomic parasites (the only homolog that appears to have entered the mainstream of cellular function is telomerase, a distinct eukaryote-specific reverse transcriptase involved in maintaining the integrity of chromosome ends). Then, finally, two connections were made (Fig. 4.5). First, the structure of the catalytic domain of one type of eukaryotic adenylate cyclase was solved, and it had the same palm topology as the RNA-dependent polymerases. In fact, the reaction of nucleic acid polymerization is mechanistically similar to the formation of a cyclic nucleotide: In both cases, the 5' phosphate is attached to the 3' hydroxyl, and the difference is whether these two groups are in two different molecules, as in polymerization reaction, or in the same molecule, as in cyclic nucleotide synthesis. Second, sequence similarity between the same adenylate cyclase and a large, mysterious family of bacterial proteins, known as the GGDEF family after the most conserved peptide, was shown (Pei and Grishin, 2001), suggesting a palmlike fold and a role in nucleotide conversion for the GGDEF family. Both predictions were confirmed; the fold of GGDEF proteins is similar to the palm domain (and GDE tripeptide is homologous to xDD in polymerases), and some of the GGDEFs are diguanylate cyclases. In this case, however, GGDEFs and adenylate cyclases, although distantly related to each other, are still closer than the most dissimilar virus polymerases.

I described some notable work on comparative virology from 1971 to the early 1990s. Even before virus nucleic acids were completely sequenced, they seemed small enough to be amenable to genome-level analysis. The ideas and approaches first brought up in connection with virus genome comparisons would reach the full bloom in the second half of the 1990s, when, finally, complete genomes of cellular life-forms were sequenced. However, before the completion of the first bacterial genome of *H. influenzae*, comparative genomics of RNA viruses produced a crescendo: in 1993, when Eugene Koonin at the National Center for Biotechnology Information, and Valerian Dolja, then at Texas A & M University (currently at Oregon State University), published a long article titled “Evolution and Taxonomy of Positive-Strand RNA Viruses: Implications of Comparative Analysis of Amino Acid Sequences” (Koonin and Dolja, 1993). Not only does this work demonstrate the power of comparative analysis for understanding ancient events in virus evolution but also it previews the developments in computational genomics of the cellular organisms, which will be examined in the rest of this book. Here, I list several themes that take us from here to there:

1. Weak similarities between viral proteins are important; they take the form of conserved sequence motifs, which can be validated by comparison of sequences with known properties and by other auxiliary information, such as similar genomic layout. If analyzed correctly, motifs reveal the mode of sequence evolution, where signals indicative of homology, common
DNA polymerase I (phages, eukaryotes, and subset of bacteria)

**DPOLI_Taq**
601 EKDWLGALVQSLQELRVLHLK 125 AFNPQVPDQATDMDENKLKMLKFLPRLL-

**DPOLI_SstI**
348 ESDWLIFANQYQQELALVHLK 125 AMMPFXQGADAILKMARLIRALK-

**DPOLI_EcoL**
375 EHPDYSIALRQYDQELALVHLK 125 AINQFADYKVRDQADYDNLAMDLWQL-

**DPOL_T7**
460 GKPWQVAGTASGLELCLHLP 110 ALTNLGSAALGSLKIWTTEMLV-

**DPOL_Yqor**
395 GLWENYTVDSFLPSLHILTH 87 WCKACRASLVTGPQGTRRIK-

**DPOL_Sce**
585 LYNVPFLVLQSQSLFQVQY 84 YAKRNLVNSLQPSLIPFQFQ-

**DNA polase, soluble (bacteria)**

**ACYC_Tbu**
897 TDPVLPLPSQTESTALAWAHP 0 ---DLMDPVAHHAVHMRSSL---

**CC1415_Ccre**
405 RAMKANPAPQVKAQALQQQ 0 ---PPVDFQGNYFTLaare-

**ACYC9_Hsap**
390 IEVVIFLAPFVTENNSHA 0 ---ANAALNLLHLMFFRLCL-

**Cyc_Amph**
960 PLTLVLFSRQPMVTTALHR 0 ---SRVAKLNNKLLFTMK---

**mil10576_Mlot**
35 RKVLTLACVIWASTELGALL 0 ---TEDFRLILAPFMQAAKAI---

**OGyc____EcoL**
889 FDQVTYFSELQVGTIPSSL 0 ---PILYVGLSLNITYFMDVQ-

**Predicted polymerases involved in DNA repair or in small DNA pathways (bacteria and archaea)**

**PH0162_Phor**
430 AKLQVQKGRPHLGLPQ---S 4 IEYAFASPRMDFPYFPHFLQKIR---

**MJ1672_Mija**
577 TRKIGLESDKNQLGPTGGL 5 IRRMTLSLMSLEFTGYYMPLL--

**TM1818_Tmar**
488 GKKLLASSLVDKLEKPGKL 4 LRVFSTLRDFKSFERKLEKVE---

**TM1794_Tmar**
493 NYGIALLCQNKDNGDWMGLGE 35 PHFAKSGIRTHSLGFQGLK---

**AP1657_Aful**
697 PRTKYAAGLQEKDGDLKQE--- 33 PAANHSSLAKENPSVNPVVP---

**HN0328_Bhal**
326 TYYAFVSLQDGNKQKL--- 4 IHNQAFQSKLXSEPAKARKNNVT---

**aq_387_Aasa**
339 MVFQIFLVMQVQVQVLQGLNL--- 10 FNEFLAVFMAELQDGLKXGLGLISHLLEKLQP---

**SBOI429_Scol**
13 SRHLYLGLDFGQVPSNGAKFG--- 16 FSFEYDSVFDFDGQKVSRKMYTL---

**Viral DNA-dependent RNA polymerases**

**RDRP_HC**
2630 XRNPGPKSYRCCPSFTFEDYND 51 VLTSCGNTLCYKLASSACAAA---

**RDRP_PV**
1972 LMEKELQFAYFQGDSALHPFB 23 NCSTSFNEMMSSLNLRTPEVTYYK---

**RDRP_Phi6**
315 KWSCLCTTMCEDHDTFSPGWNL 58 GQATDMGQVYLSNAAVWGVQDRTAP---

**Reverse transcriptases - viral and cellular (telomerase)**

**RT_MMLV**
315 PHQVYSTQXKAPDCCFCLRLP 26 QGQFQMVADDPEARLDRADIFR---

**RT_NIV**
101 KKKKQVTVLVGADFYYFVYNF 27 PQNQGKPSAIFQSMKIERFFEPKL---

**RTF_HsaP**
581 PELETFYVAKVHYDSDGYKQD 107 PQGSLHSTLSLSGYGDMRNLHAP-----

**TRP_Spm**
581 GKRYTFRVRKSSCQMRKLQD 107 PQGSLHSTLSLSGYGDMRNLHAP-----

_**Figure 4.5.**_ Conserved sequence motifs in viral RNA-dependent RNA polymerases, viral and cellular reverse transcriptases, and their cellular homologs with polymerase and nucleotide cyclase activities. Modified from Makarova et al., Nucleic Acids Research Vol. 30, 2002 by permission of Oxford University Press.]
function, and similar structure persist in the form of short regions of conservation interspersed with nonconserved regions of variable length. This mode of evolution is relevant not only to virus proteins but also to any large and diverse superfamily of proteins in prokaryotes and eukaryotes.

2. Although the diversity of viruses may be mind-boggling, the number of building blocks—in this case, conserved virus proteins—is limited. Only one protein, RNA-dependent RNA polymerase, is found in all nondefective RNA viruses; nevertheless, the majority of virus genes are conserved in several virus groups, and none of the viruses contain more than 20 discrete genes. In Chapter 5, we will see that the number of gene/domain blocks that make up the cellular organisms is also finite, although of course much larger than in viruses, and is amenable to analysis by present-day computer technologies.

3. Clustering by similarity produces a limited, tractable number of related protein groups (families or superfamilies; there is no real difference between the two). For example, all RdRp enzymes in RNA viruses fall into one of three conserved superfamilies. Likewise, all RNA helicases encoded by positive-strand RNA viruses belong to one of the three helicase families. Chapters 5 and 10 discuss how proteins encoded by complete genomes can be sorted into a finite number of protein families, greatly facilitating the analysis of the molecular setup of a cell.

4. Combinations of genes are not random. Three types of RdRps and three types of helicases could give rise to nine polymerase–helicase combinations; however, in nature, only three such combinations are widely distributed. Chapter 7 discusses the tendency of some genes to occur in conserved arrays.

5. Conserved arrays of genes or protein domains are not merely the auxiliary factor in determining the identity of protein sequences. The order of genes in virus genomes is a molecular trait in its own right, as indeed is the information on gene co-occurrences mentioned previously. These traits can be used for function prediction and in evolutionary reconstruction.

6. There is abundant evidence for recombination between different virus genomes. Extended regions of nucleotide identity are apparently not required for such recombination. Recombination complicates reconstruction of phylogeny, but evolutionary and taxonomic chaos does not ensue, as discussed in Chapters 11 and 12.