Genomic signatures in viruses

Joel Gustafsson

Department of Infectious Diseases

Institute of Biomedicine

Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg, Sweden, 2023

Genomic signatures in viruses

ISBN: 978-91-8069-269-4 (PRINT)
ISBN: 978-91-8069-270-0 (PDF)

Printed in Borås, Sweden 2023
Printed by Stema Specialtryck AB

"Nothing in life is to be feared, it is only to be understood.
Now is the time to understand more so that we may fear less."
-Marie Curie

Abstract

In an age of global pandemics, studying how viruses and their genomes evolve is of great importance. It has previously been found that genomes of many eukaryotes and prokaryotes have specific preferences for nucleotides, dinucleotides and codons. Such preferences are characterized by the selective pressure acting on the genomes, and are referred to as specific genomic signatures. The presence of such signatures has, to our knowledge, not been studied in viruses, and it is therefore the aim of this thesis to thoroughly investigate genomic signatures in viruses.

In the first two papers of this thesis, new algorithms for the study of genomic signatures were developed. Here, such genomic signatures were based on variable-length Markov chains of a genome. Compared to pre-existing methods, our new algorithms are a thousand times faster, and compared to the state-of-the-art, the algorithms are up to 600 times faster while also requiring less memory. These methods enable computationally efficient analysis of genomic signatures, even on laptops.

In the subsequent two papers, we thoroughly analyzed the genomic signatures of viruses and compared such signatures to those of the viruses' hosts. The results illustrate that a majority of viruses have specific genomic signatures. In addition, in most cases, the signatures of viruses are not similar to the signatures of their hosts other than in GC content. This dissimilarity indicates that viruses' signatures are independent of their host's signature, despite viruses' dependence on their host's genetic and protein-expression machinery.

In the final paper, we illustrated an application of the genomic signatures by applying them to identify recombination events between *Human alphaherpesvirus 1* and *Human alphaherpesvirus 2*.

We thus demonstrate that genomic signatures of variable length are an important property of virus genomes. They hint at the importance of the evolution of specific patterns of the nucleotide sequence of viruses. These patterns can likely identify even remotely related viruses in collections of unknown sequences, thus helping detect and classify novel viruses. In addition, it might be possible to use and modify the genomic signatures to, e.g., attenuate viruses to create vaccine candidates.

Keywords

Virus evolution, Virology, Bioinformatics, Markov Chains, Alignment-free

Sammanfattning på svenska

För att kunna undvika framtida viruspandemier krävs en djupare insikt i virusevolution. Tidigare studier har visat att i eukaryoters (t.ex. djur och växter) och prokaryoters (t.ex. bakterier) genomen finns det en specifik preferens för vissa nukleotider, dinukleotider och kodon. Dessa specifika preferenser i en organisms genom kallas för dess genomiska signatur och formas av de selektionstryck som har inverkan på genomet. Det är denna avhandlings syfte att undersöka om även virus har specifika genomiska signaturer.

I avhandlingens första två manuskript utvecklade vi nya algoritmer och verktyg för att kunna genomföra studier av virus genomiska signaturer. De genomiska signaturerna modelleras här av Markovkedjor med variabel längd, för vilka de tidigare tillgängliga algoritmerna är antingen långsamma eller kräver mycket minne. I jämförelse med den tidigare snabbaste algoritmer är vår implementation upp till 600 gånger snabbare samt kräver betydligt mindre minne. Detta gör att genomiska signaturer kan analyseras på samtliga genom och med lättillgänglig hårdvara, till exempel bärbara datorer.

I följande två manuskript genomförde vi en noggrann undersökning av vilka virus som har specifika genomiska signaturer, och om de är art-, genus-, eller familjespecifika. Resultaten illustrerar att i många fall har virus specifika signaturer. Dessa specifika signaturer är bevarade i hela genomen och i vissa fall väldigt lika signaturer från närbesläktade virus. Dessutom visade vi på att dessa signaturer i många fall är skilda från signaturerna hos virusens värdar, vilket är anmärkningsvärt eftersom mycket av de selektionstryck som verkar på virusets genom tros vara liknande de som verkar på värdens genom.

I det sista manuskriptet visade vi på ett användningsområde av genomiska signaturer, för detektion av rekombination mellan två arter av virus.

Vi har visat att virus har specifika genomiska signaturer som skiljer sig från deras värdars genomiska signaturer. Detta antyder att det finns viktiga evolutionära fördelar för virusen att forma deras genom på vissa specifika sätt. Våra insikter kan eventuellt användas för att klassificera sekvenser av okänt ursprung samt för att förändra virus genom i syfte att tillverka vaccin.

List of papers

This thesis is based on the following studies, referred to in the text by their Roman numerals.

I Gustafsson, J., Norberg, P., Qvick-Wester, J.R., Schliep, A.
Fast parallel construction of variable-length Markov chains
BMC Bioinformatics 22, 487 (2021)

II Gustafsson, J., Edwards, S. V., Schliep, A., Norberg, P.,
Estimating phylogenies from raw sequencing reads using variable-length Markov chains
Manuscript

III Holmudden, M.*, Gustafsson, J.*, Schliep, A., Norberg, P.
Species-specific genomic signatures in viruses
Manuscript
* Denotes shared first-authorship

IV Gustafsson, J., Schliep, A., Norberg, P.
Virus-host similarities in genomic signatures
Manuscript

V Gustafsson, J., Schliep, A., Norberg, P.
Detection of Herpes simplex type 1 and 2 recombination in clinical samples
Manuscript

Contents

Chapter 1 Introduction

1.1 Viruses

Viruses are small ubiquitous parasitic agents that cannot replicate outside of the cells of their hosts. They are believed to infect all domains of life and are thus not only important human pathogens, but also major contributors to most, if not all, ecosystems on the planet. Upon infection of a host cell, viruses use parts of their host's cellular machinery to produce proteins and reproduce. Despite these common themes, there is an enormous diversity in the viral domain.

As an example of this diversity, consider the different mechanisms involved in the replication of viruses. Some viruses are entirely dependent on their host's replication machinery to copy their genomes, such as *Papillomaviridae* [1]. Other viruses contain genes that facilitate the replication of their genomes. For example, many viruses with RNA genomes, such as the Poliovirus [2], contain an RNA-dependent RNA polymerase (RdRp), and some DNA viruses contain other proteins that facilitate replication, such as the Herpes simplex viruses [3], and Poxviruses [4]. Likewise, most viruses rely on their host cell's genetic machinery to express their genome and create proteins, although some viruses also contain genes that participate in this process [5].

After the virus has replicated, the virus can leave its host cell through various mechanisms. These mechanisms include budding, where the virus particle leaves the cell by taking part of the host cell's membrane, common among *Herpesviridae* [6]. Additionally, viruses can use apoptosis—where the host's cell is lysed, and virus particles are released, e.g., *Adenoviridae* [7]—or exocytosis—where multiple virus particles are released from the cell with the help of the host's transport system, e.g., *Picornaviridae* [8].

Some viruses wait a long time before leaving their host cell. For example, some viruses that infect bacteria, such as the lambda phage [9], integrate into their host's genome and are replicated along with the host's genome as the host cell divides. Other viruses, such as the human immunodeficiency viruses [10], integrate into their host's genome, where they can remain dormant. Other viruses do not integrate into the host's genomes but still lie dormant in specific target cells with typically limited symptoms, e.g., Herpesviruses [11].

The first virus discovered was the plant pathogen *Tobacco mosaic virus* in 1892 [12], although the term 'virus' was first used in 1898 [13]. Today, it is estimated that viruses

are the most abundant biological entity on earth, with at least one viral species per other species [14]. Despite this abundance, the International Committee on Taxonomy of Viruses (ICTV), which organizes the taxonomy of viruses, only recognizes 10 434 virus species in 2022 [15].

1.1.1 Structure of viruses

Viruses protect and transport their genomes from one host to the next in virus particles (virions) which is typically composed of an outer shell (capsid) and sometimes also a lipid layer (envelope). The capsid is typically composed of a single or a few small proteins, and the most common shape is an icosahedron, a 20-sided structure that closely resembles a sphere [16]. For some viruses, the capsid is additionally surrounded by an envelope. This envelope is composed of lipids and is usually acquired from one of the host cell's membranes. In addition to protecting the genomic material, the capsid and envelope are instrumental in allowing the virus to enter new host cells.

1.1.2 Genomes of viruses

In the viral domain, different viruses use different types of genomic material for their genomes. The genomic material is composed of either DNA or RNA, which can be double or single-stranded, with either orientation of the single strand. In total, seven types are recognized and organized into the Baltimore classification based on the method of mRNA synthesis [17, 18]: (i) double-stranded DNA, (ii) single-stranded DNA, (iii) double-stranded RNA, (iv) positive single-stranded RNA, (v) negative single-stranded RNA, (vi) retro-transcribed single-stranded RNA, and (vii) retro-transcribed double-stranded DNA.

Furthermore, some viruses have genomes that are composed of several parts, referred to as segmented genomes. These segmented viruses include, for example, the Influenza viruses.

Virus genomes are, on average, small in comparison to the genomes of eukaryotes and prokaryotes. Nevertheless, there is a large range of genome sizes among these small genomes. For example, the *Porcine circovirus* has a genome that is only 1729 nucleotides long [19], while the current largest known virus genome belongs to the *Pandoravirus salinus* which has a genome of 2.5 million nucleotides [20]. The latter rivals the size of bacterial genomes (Figure 1.2).

Perhaps due to how short viral genomes are, most viral genomes are tightly packed with genes, with an average of 6 % to 10 % of their genome not coding for genes [21]. In prokaryotic genomes, 6 % to 25 % of the genome is non-coding [22, 23], and for eukaryotes, as much as 98 % of the genome can be non-coding [24]. Nonetheless, there is still a considerable difference between the sizes of coding virus genomes and coding prokaryotic and eukaryotic genomes, both in terms of the number of genes and the length of the coding sequence.

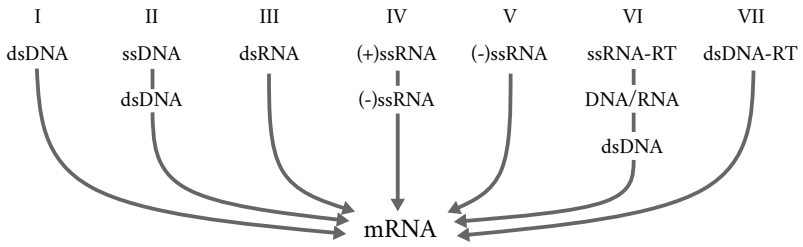

Figure 1.1 Illustration of the Baltimore classes. The Baltimore classification is based on the method of mRNA synthesis. Highlighted here are the intermediary products for the different types of genome compositions in the viral domain.

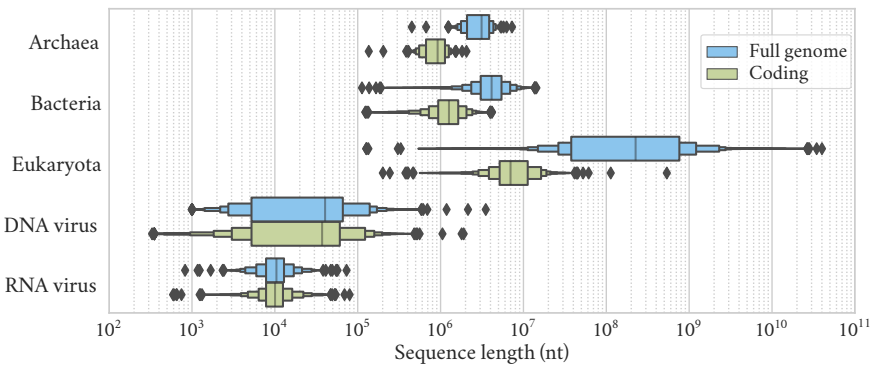

Figure 1.2 The range of genome sizes. Includes many bacteria, viruses, and eukaryotes available in NCBI.

1.2 Virus evolution

Viruses are under constant pressure to evolve to overcome evolutionary barriers [25]. Such barriers include their host's virus defense or the specific cellular environment. Organisms are able to adapt to their environments by processes acting on random changes (mutations) in their genomes. Variants of the same genome or gene with different mutations are referred to as alleles. While viruses are here used as examples throughout, the concepts described apply to the evolution of all organisms.

1.2.1 Natural selection

Natural selection describes a primary method of evolution [26, 27]. According to natural selection, alleles that increase a virus' reproductive success will, over time, become dominant in the population as those individuals can out-compete others. Such beneficial alleles are generally said to increase the fitness of the individual or population. Examples of alleles that increase the fitness of viruses could be alleles that help a virus infect new cells or those that help it escape the host's immune response.

As an example of natural selection, consider a genetic variant of a virus (a virus strain) with an allele that allows it to produce five new viruses per infected cell, while the strains without this allele only produce two new viruses. After the first generation of these viruses, there are three more strains with the beneficial allele. After the next generation, there are 25 viruses with the beneficial strain and only four of the non-mutated strain (see Figure 1.3 for an illustration). Over several generations, the strains with the beneficial allele will dominate the entire population. Likewise, strains with alleles that have a negative impact on the virus will lead to fewer descendants, and thus those strains are likely to be out-competed by other strains.

Natural selection is an important evolutionary mechanism for viruses as they are constantly evolving to compete in their environment. Viruses compete not just with other viruses for resources but also with their hosts. The evolution of viruses is complex, and an allele that increases the fitness of a virus in one host or environment may be detrimental in another host, for example, due to differences in the adaptive immune system.

1.2.2 Genetic drift

Genetic drift is an additional evolutionary process that acts on all alleles, regardless of their impact on the fitness of the population [28]. Genetic drift is based on the random events that influence which alleles are carried over from one population to the next, for example based on which viruses successfully infect a new host. Due to this randomness, the frequency of alleles in a population can change over time and, in some cases, lead to significant variation in which alleles are present in a population. This process can be particularly notable when two populations are physically separated, such as the infection of two hosts by the same virus strain. Despite being subjected to similar selective pressures, the frequency of alleles among the viruses in the two hosts can be slightly different.

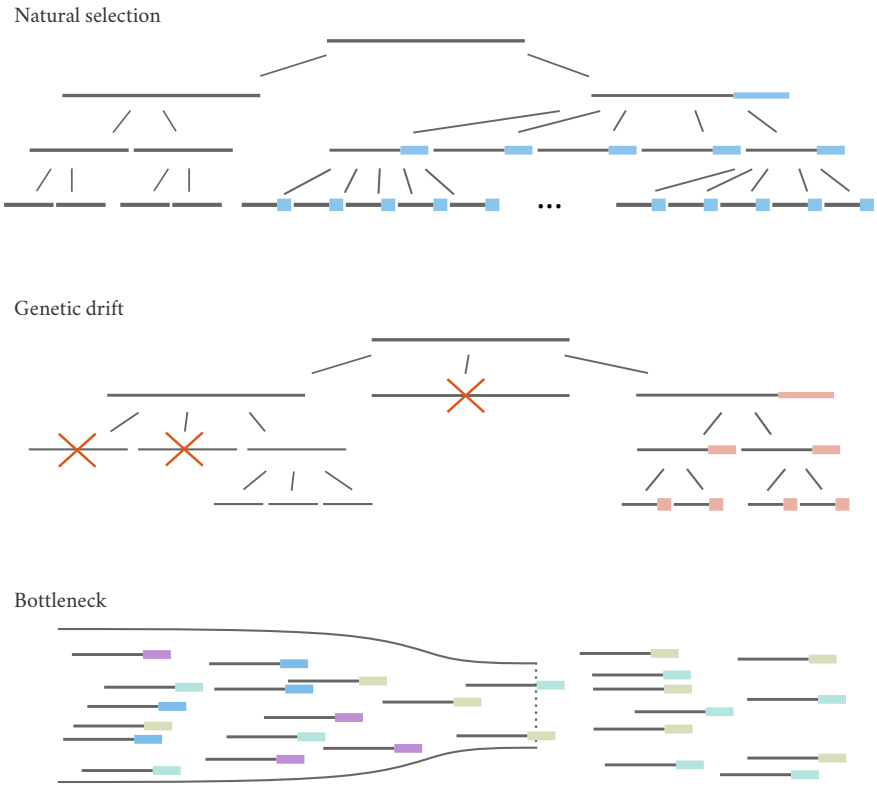

Figure 1.3 Illustration of three evolutionary processes. In the natural selection example, the blue allele (thicker line) replicates more efficiently and eventually dominates the population. In the genetic drift example, the red allele (thicker line) replicates less efficiently, but due to random effects, some strains do not successfully replicate. In the bottleneck example, only part of the population survives (the lighter green and blue colors), which influences the alleles of the surviving population.

The effect of genetic drift is more notable in smaller populations or when only a few viruses carry a specific allele. The randomness in reproduction has a larger influence on smaller populations as there is a greater possibility that only viruses with a certain allele randomly do not reproduce. Specifically, this process can lead to alleles with a negative influence on the fitness of a population becoming prominent [29].

As an example of genetic drift, consider three viruses that infect three cells. One of these viruses has a negative allele that causes it to reproduce slower than the other two. Randomly, one of the cells infected with one of the faster viruses dies. In the other two cells, the viruses can replicate and produce three and two progeny viruses, respectively. In the next generation, two of the viruses that replicate normally do not reproduce, while the others do, leading to three normal viruses and a total of four viruses that reproduce slower. Despite producing fewer progeny viruses, the fraction of viruses with the negative allele has increased from 33 % to 57 % (Figure 1.3).

1.2.3 The bottleneck effect

Bottleneck events are an important driver of genetic drift [29]. In a bottleneck event, only a small portion of the original population can reproduce. Examples of bottleneck events include those where a large portion dies out, either due to random events or external factors. Alternatively, the surviving population can have some property that enabled it to survive this specific event without being generally advantageous. This is similar to the founder effect, which describes that when a subset of a population moves to a new area, the resulting population will be dominated by the alleles in that subset. The alleles in the population after the bottleneck can be neutral or even disadvantageous in the environment (see Figure 1.3 for an illustration).

One prominent bottleneck event for viruses is the transmission of the virus from host to host [30]. For example, the alleles that enable an optimal expression of a virus' genome and replication of the genome in the host cell may not be optimal as the virus particle is transferred from one host to another. Likewise, none of the alleles that do not permit the virus to spread from one host to the next will be present as the virus infects a new host.

1.2.4 Speciation

New virus species can be formed in various scenarios. Among them are co-speciation, intrahost speciation, and speciation from host switches.

Co-speciation refers to when viruses speciate as their host speciates. For example, many herpesviruses have speciated along with new species of their hosts [31]. Under this mode of speciation, we might expect that the rate of virus evolution is consistent over time as the viruses gradually adapt to an evolving cellular environment that is relatively similar to the previous host.

Intrahost speciation is the speciation of viruses where a virus species is duplicated into two different species that infect the same host. For example, intrahost speciation has been suggested to have resulted in the virus species *Equid herpesvirus* 1, 3, 4, and 8 [32]. Under this mode of speciation, the virus species likely must have limited interactions between their genomes, as they otherwise might recombine to the same strain. Such recombination can be restricted, for example, by occupying separate niches in the host. It has also been suggested that some herpesviruses prevent recombination through the creation of separate replication compartments [33]. Another method of preventing recombination might be to drastically alter the nucleotide sequence to prevent the initiation of homologous recombination [34].

Speciation due to host switches occurs as a virus strain evolves to infect a new host species. This is called zoonosis if the new or old host is human. For example, this is the origin of HIV-1 and HIV-2, with a host switch from chimpanzees and Sooty mangabeys, respectively (eviewed in [35]). As the virus adapts to the new host, it is typically subject to a different selection pressure than in the previous host. This altered selection pressure can lead to a significant amount of mutations and, over time, a significantly different virus genome.

1.2.5 Molecular processes underlying evolution

Single nucleotide mutations

The mutations that enable viral evolution stem from processes that introduce changes in the genome. Such changes can, for example, be introduced during genome replication, from the host's immune system, or stem from damage caused by external sources. The mutation rate of viruses is typically orders of magnitude higher than their hosts', but the rate varies considerably depending on the virus and genome type [25]. For example, the *Tobacco mosaic virus* is estimated to mutate with a frequency in the range of 7.26 to 10.3×10^{-6} [36], while the human norovirus has an estimated mutation frequency of 1.5×10^{-4} [37]. However, recently, it has been found that the long-term evolutionary rate of viruses might be closer to the host's evolutionary rate, based on the observation that over longer time periods, many of the introduced mutations are reversed [38, 39].

Gene duplication

Gene duplication is an important evolutionary process that enables the generation of novel genetic material in genomes [40]. There are several potential outcomes of gene duplication. The genetic sequence can be preserved for redundancy or to increase the expression of the protein encoded by the gene. Alternatively, because there is an intact copy of the gene, the duplicated gene is free from selective pressure and can be mutated to generate novel gene functions.

Despite being a common occurrence in prokaryotes and eukaryotes [40], only a small number of genes are reported as duplicated in RNA viruses [41]. In contrast, gene duplication is common for dsDNA viruses but infrequent for ssDNA viruses [42]. For example, for the *Acanthamoeba polyphaga mimivirus* between $26.3\,\%$ and $35.0\,\%$ of the genes have a homologous gene in the genome, likely caused by gene duplication [43].

Recombination

Another feature of viral evolution is that of genetic recombination [44]. In a recombination event, genomic material is exchanged between two genomes, often dependent on identical inherited (homologous) sequences in the two genomes. For many viruses, recombination serves two purposes: (i) to gain several beneficial mutations and (ii) to expel harmful mutations.

The presence of recombination in viruses has a beneficial impact on their evolution [44]. It allows beneficial mutations from different strains to combine in a single genome and thus form an evolutionary more fit individual genome. For example, recombination between different species is believed to enable host-switches of, e.g., *Coronaviridae* [45, 46], *Geminiviridae* [47], and *Papillomaviridae* [48].

In the absence of recombination, according to Muller's ratchet [49], mutations with negative impacts (deleterious mutations) accumulate in a population's genomes. This effect stems from the observation that the number of deleterious mutations likely

outnumbers the beneficial mutations. Thus, recombination is crucial as it allows deleterious mutations to be removed from a genome by exchanging the region with such mutations with a region without the deleterious mutations.

Reassortment

Segmented viruses are subjected to an additional evolutionary mechanism. When multiple segmented viruses simultaneously infect the same host cell, new variants can form by combining segments from different strains. This enables beneficial mutations in one strain to be combined with beneficial mutations in a different strain. This is exemplified by the influenza A viruses, where this process has led to human pandemics [e.g., reviewed in 50].

1.2.6 Virus-host evolutionary relationship

Viruses are under constant evolutionary pressure to adapt to their evolving host and its immune system [39]. This is referred to as the "Red Queen hypothesis", in reference to viruses needing to constantly evolve to maintain their fitness in their environment, as the host attempts to rid itself of the virus [51].

Most studied virus-host relationships are pathogenic, where the virus is harmful to its host. However, not all viruses are *only* harmful to their hosts [52]. Notably, some viruses can benefit their host and make the host more competitive in its environment, for instance in some bacteria, by infecting and killing competitors while providing immunity to the same infection [53].

The apparent symbiotic human viruses were recently reviewed [54]. A number of viruses were identified as symbiotic in the sense that they are not necessarily associated with any disease. Such human viruses include those that help the immune system develop at an early age (e.g., Pegiviruses) and those that protect from infection by other viruses (e.g., Anelloviruses).

There are many reasons for the adaptation of the viral genome towards the virus' specific host environment. Due to the high mutation rate, it is reasonable to assume that viral genomes are highly optimized for their current host or hosts. For viruses that infect many hosts (generalist viruses), such adaptation may not focus on a specific host but on something that works in all hosts (e.g., reviewed in plant viruses [55]).

Additionally, many viruses that have infected humans for a long time have evolved to cause relatively mild symptoms [56]. This is based on the assumption that if a virus causes less severe symptoms, its host is more prone to meeting other potential hosts and thus spread the virus further. If, instead, the virus quickly kills the host, it only has a small amount of time to spread to a new host. However, it is not known if viruses consistently mutate to decrease their virulence, and they could also become more virulent.

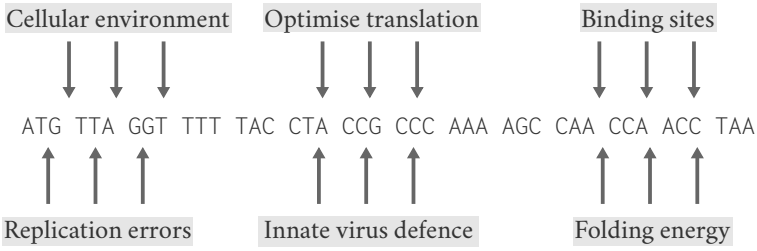

Figure 1.4 Illustration of possible selection pressures acting on genomes. Genomes are adapted based on their cellular environment and through, for example, replication errors, to, e.g., have an optimal translation and folding energy, and to not be targeted by anti-virus systems in the hosts.

1.3 Nucleotide sequence adaptation

A significant portion of the evolution of virus genomes relates to the amino acid sequence and the resulting proteins. Such protein evolution includes increased binding affinity to a host's receptors or avoidance of the human adaptive immune system by altering the conformation of the surface proteins. While most studies of adaptation in viruses have focused on protein-coding adaptations, it has also been suggested that viruses also adapt their nucleotide sequence to their environment [57]. Nucleotide adaptation is possible without influencing the amino acids due to the redundancy of codons. For each amino acid, there are between one and six possible codons that are translated to that amino acid. Typically, the last of the three nucleotides of the codon can be altered without changing the amino acid. Changing the nucleotide sequence without altering the amino acids is complicated by the presence of overlapping reading frames, where the same sequence codes for multiple proteins by initiating transcription at multiple points [58, 59]. It might be possible to adapt the nucleotide sequence further by mutating to an amino acid that has similar properties to the original amino acid.

1.3.1 GC content

One of the most general properties of a genome is its nucleotide composition. The nucleotide composition is often reduced to the GC content, defined as the fraction of the genome that is either `G`s or `C`s. The GC content of a genome is an important factor for the thermal stability of the genome, with a larger fraction of `G`s or `C`s providing higher stability [60]. However, there is no significant correlation among prokaryotes between the average GC content and optimal growth temperature [61]. Instead, prokaryotic communities appear to prefer different GC contents depending on the environment, with different averages of GC content for microbes in soil as compared to ocean samples [62]. In other organisms' genomes, for example in the human genome, there is significant variation in GC content [63], where regions with low GC content typically have a lower gene density [64, 65].

1.3.2 Dinucleotides

In addition to GC content, many adjacent pairs of nucleotides (dinucleotides) occur with a frequency not expected from the nucleotide content. For example, the dinucleotide CpG can be significantly less common than expected from the number of Cs and Gs. That some dinucleotides, most prominently the CpG dinucleotide, are less frequent than expected in the genomes of vertebrates has been noted for a long time [66]. This is primarily thought to be linked to the epigenetic-related methylation of the CpG dinucleotide. When the methylated CpG dinucleotide is spontaneously deaminated, the dinucleotide TpG is sometimes formed and not always properly corrected [67].

Some viral genomes have been shown to mimic their host's dinucleotide usage, in particular, to avoid the CpG dinucleotide [68, 69]. This dinucleotide is targeted by, for instance, the APOBEC and ZAP families of proteins in vertebrates which deaminates the CpG dinucleotides, leading to a decreased fitness of the virus [70–73].

The dinucleotide usage has been illustrated to be unique to some genomes [74]. In addition, this specific preference for certain dinucleotides was illustrated in *E. coli* and *C. elegans* to be conserved throughout most of the genome [74]. This indicated that by analyzing the dinucleotide content of a sequence of DNA, it would be possible to identify which genome that sequence belongs to. Based on this observation, the measure of dinucleotides could be used as a "genomic signature". This concept of a genomic signature is the basis of this thesis, although we will expand it to include more than dinucleotides.

This genomic signature based on dinucleotide frequency was further analyzed in a larger set of genomes [75], where the specificity of the signature was further cemented. There, the authors compared the average dissimilarity of dinucleotides in 50 kbp samples of a set of eukaryotic and prokaryotic organisms. Among these 50 kbp samples, they found that the average dissimilarity in a species was significantly smaller than the average dissimilarity between other species.

1.3.3 Codon adaptation

The genomes of many organisms are also biased towards the usage of specific codons, referred to as codon usage bias [76, 77]. One of the major determinants of which codons are used to express a given amino acid is believed to be related to the specific genetic machinery of the organism [78]. In particular, during the translation of a gene, a specific set of transfer RNAs (tRNAs) are typically more abundant, making the expression of genes that utilize those tRNAs more efficient. Therefore, there might be selective pressure on the genome to use that set of tRNAs [79, 80]. This codon adaptation is thought to be especially important for highly expressed genes.

One of the major contributions to codon usage is that of the GC content of the genome. This correlation is particularly notable for bacteria, where codon usage can be estimated from the GC content of non-coding regions [81]. Likewise, in some DNA viruses, the codon usage is determined by the GC content [82]. In some other organisms, there is a considerable variation between genes in which codons are preferred [83]. For instance, human genes have a large variation in the preferred codon usage [83]. Although, this has partly been explained by tissue-specific codon usage [84].

1.3.4 Codon pair bias

In addition to the codon usage being biased in some organisms, the codon pairs appear in a frequency not expected by the codon usage [85]. The specific mechanisms that are influenced by the codon pair bias are not yet elucidated, but in yeast, the codon pairs appear to influence the efficiency of the translation machinery [86]. This impact on the translational efficiency was suggested to be caused by interactions between adjacent tRNAs during elongation. Furthermore, codon pairs have also been illustrated to have a significant influence on mRNA stability [87]. Specifically, genes with an altered codon pair usage were expressed in cells, and the altered genes were illustrated to have a less stable mRNA than the wild-type gene. This codon pair bias has, e.g., been used to attenuate a polio virus by changing the frequency of codon pairs [88].

1.4 Alignment-based sequence comparison

Bioinformatics deals with computational methods of analyzing and interpreting biological data. Already in the 1960s, computers were used to analyze sequences and their evolution based on sequence similarity [89]. Today, the size of the problems has grown along with the myriad of data types and methods available. However, the identification of sequence similarity is still a foundational problem, and is computed through sequence alignment.

A sequence alignment of two or more genomes is an arrangement of the sequences which indicates similarities (matches between the sequences) and dissimilarities (mismatches, substitutions, and deletions). See Figure 1.5 for an example.

Figure 1.5 Illustration of an alignment. Here '|' indicates a match between the two sequences, '-' indicates an insertion or deletion, and unmarked positions indicate other mismatches. Alignments derive their importance from the fact that an alignment with many matches hints at the existence of a common ancestor. **s** and **t** likely share a common ancestor, as they have much shared sequence.

There are many ways to formalize what constitutes the best alignment and many ways to compute the best alignment given any specific mathematical definition by assigning a gap penalty and a score for pairing matches or mismatches, e.g., the Dayhoff's famous PAM matrices [90]. The resulting alignment depends on the choice of parameters [91]. In some cases, such as when the assigned penalty for mismatches far exceeds that of gaps, the alignments end up consisting primarily of gaps. An early alignment algorithm is the Needleman-Wunsch algorithm [92], which calculates global alignment, i.e., the alignment of entire sequences. Aligning sequences is computationally expensive, requiring time in the order of $n \times m$ with n and m being the respective length of the aligned sequences. This makes the task computationally prohibitive for longer

sequences, such as chromosomes or even the largest genes (in the human genome 2.4 Mbp [93]).

In many applications, such as estimating the evolutionary history of a collection of sequences, an alignment of all sequences in question is needed, referred to as a multiple-sequence alignment. This makes the problem significantly harder as the alignment of each sequence needs to consider every other sequence. In addition, this adds to the computational complexity. In fact, there likely is no computationally efficient algorithm that solves the problem exactly [94]. Instead, many methods resort to approximations.

Perhaps the most popular alignment-based tool is `BLAST`, which uses short exact matches in sequences to gradually create a local alignment of a single sequence to a large collection of sequences [95]. A local alignment is defined as an alignment of a subsequence of the first sequence to a subsequence of the other sequence. This is especially useful when the two sequences in question are dissimilar or where parts of the sequences are highly mutated while others are conserved.

Sequence alignment is an invaluable tool for many biological tasks, such as finding the origin of a newly discovered sequence and determining the evolutionary relationship between sequences. This is based on the observation that if two sequences align with few mismatches (high alignment score), there likely is a recent common ancestor of the two sequences (e.g., Figure 1.5).

Another important application of sequence alignment is that of read mapping. For example, given a collection of DNA sequencing reads from a human, a read mapper places each read at a genomic position in the human genome. An example use case is the detection of mutations in a genome which can indicate disease or potential treatments. Popular read mappers include `minimap 2` [96], and `Bowtie 2` [97].

1.4.1 Limitations of sequence alignment

Sequence alignment relies on a number of assumptions about the evolutionary relationships of the sequences [98]. For viruses, two assumptions are especially hard to support. (i) The alignment of sequences in question needs to be largely homologous. However, many viruses have high mutation rates [99]. This can make alignments of more distantly related or not conserved regions highly gapped and unreliable. (ii) The alignment of whole genomes assumes that the genomes are clonally related. In viruses, this is frequently violated as parts of the genomes can move between different strains, for example, through recombination. This makes estimating the evolutionary distance between two sequences based on alignment difficult. Furthermore, for alignment scores as high as 60 %, remote homologs are often difficult to distinguish from unrelated sequences [100].

To address these issues, alignments are typically limited to conserved genes, with the assumption that these can be used to represent the evolutionary history of the viruses in question. Alternatively, alignments are based on the amino acid sequence. Genes are typically selected based on their presumed shared evolutionary origin. However, the alignment of different genes from the same viruses often gives conflicting evolutionary histories, e.g., for papillomaviruses [101], and herpesviruses [102].

1.5 Alignment-free sequence comparison

To study evolution and sequence similarity without the computational demands and assumptions of sequence alignment, alignment-free sequence comparison has become a popular alternative. The most common alignment-free methods use the counts of words of length k, so-called k-mers, in sequences. For example, ACGT is a 4-mer, and CATATTAC is an 8-mer.

Early alignment-free methods considered the GC content of genomes [103], essentially equivalent to the count of 1-mers. This was later expanded to include dinucleotides (2-mers) [74], tetranucleotides (3-mers) [104], and up to 9-mers [105]. It was further found in 2003 that by relying just on GC content, it was possible to distinguish between 40 % of the then sequenced species [106]. Other early work on k-mers utilized the fact that if two sequences are homologous and align well, they must share many k-mers [107] (see Figure 1.6).

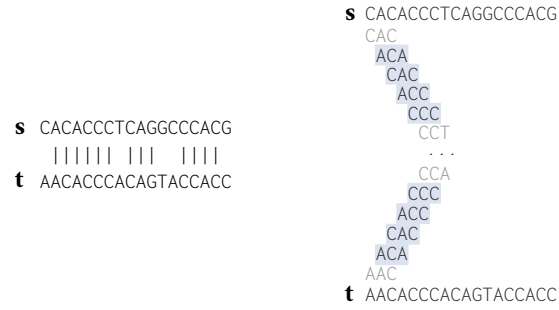

Figure 1.6 Illustration of correspondence between alignment and matching k-mers. The positions that align on the left correspond to matches between the 3-mers on the right. The blue boxes illustrate matching 3-mers and the gray 3-mers do not match.

Today, the most popular alignment-free approach considers the frequency of long k-mers (often $k \geq 20$) in two sequences and assigns a similarity score based on how many of the long k-mers are present in both sequences. This type of analysis powers many popular modern tools, such as Kraken 2 [108], which, similar to BLAST, can identify the origin of a sequence but is considerably faster. Other methods, such as Mash [109] and skmer [110] use the number of shared k-mers of two sequences to estimate their evolutionary distance.

Most long k-mers will occur approximately once in a genome (Figure 1.7). Some alignment-free methods utilize this fact to gain the advantages of sequence alignment and homology without computing the alignment. Specifically, the matching k-mers are assumed to originate from the same position in the genome. Depending on the genome size, the value of k needs to be chosen carefully to ensure that the matching k-mers come from the same location in the genome, and so that random mutations do not lead to the presence of the same k-mer at a different position. This works well for the classification of a genome and for studying the evolutionary relationship between closely related strains. However, with mutations above a certain threshold, few

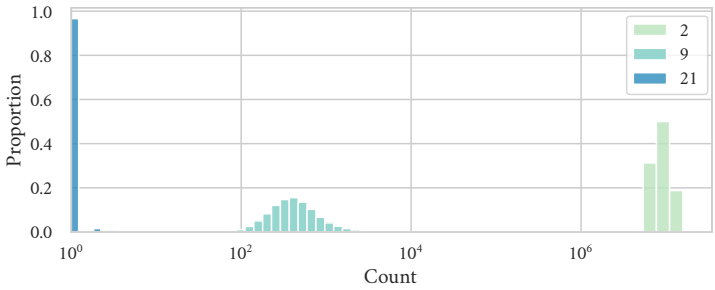

Figure 1.7 The proportion of frequency of 2,9,21-mers in *Drosophila melanogaster*. Most 21-mers occur once and 2-mers roughly one sixteenth of the genome size. For the 9-mers there is significantly more variation, where some k-mers are frequent and others are infrequent.

k-mers will be shared, which makes reasoning about the relationship of the species difficult.

Instead, the frequencies of shorter k-mers have been illustrated to be informative of both the sequence and the biology (reviewed in [34]). Here, the typical method picks a value of $k \leq 15$ and compares the count of k-mers from a collection of sequences. For example, two distantly related species might not have a significant amount of conserved sequence, but with a large fraction of inherited sequence from an ancestral species, they might still have a relatively similar frequency of many shorter k-mers as compared to unrelated species. In addition, shared frequencies of shorter k-mers can indicate evolutionary pressure acting on their genomes.

Early methods using short k-mers defined the D_2 statistic that compares exactly the count of each k-mer in two sequences to determine how related they are [111]. Later work established a background-adjusted version called D_2^* that instead compares the count of the k-mers based on the expected counts [112]. Applications of this measure include host-prediction of bacteriophages [113], based on a presumed adaptation of the virus to the host. A similar approach called `CVTree` has been used to estimate the phylogeny of prokaryotic organisms [114].

1.5.1 Limitations of *k*-mer based approaches

What the appropriate length of k is for a specific task is not obvious. When analyzing 1-mers or 2-mers, all such k-mers will exist in the sequence and be very frequent (Figure 1.7), but it might be difficult to distinguish between highly related sequences. Therefore, a slightly larger k needs to be selected, where most k-mers are still frequent but where the counts are sufficiently distinct to differentiate between related sequences. In many cases, 6-mers or 9-mers have provided a good tradeoff between these factors [113, 115]. For longer sequences, using $k = 10$ [116] or $k = 14$ [117] can give better results.

Furthermore, with the typical k-mer approaches, the size of k is fixed. This is suitable for longer genomes, where the count of each k-mer can be estimated well. However, even in longer genomes, many longer k-mers are infrequent, making esti-

mations about their frequency throughout a genome impossible. Therefore, it is often necessary to tune the size of k based on the length of the sequences. For instance, the frequency of most 9-mers in the *Drosophila melanogaster* genome varies from 100 to 5 000 occurrences (Figure 1.7). There, the estimation of the frequency of the low-count k-mers from, e.g., a fragment of the sequence, would be less accurate than the estimation of the high-count k-mers.

As the length of k increases, the number of possible k-mers in an infinitely large genome increases exponentially as $O(4^k)$. In real genomes, the number of possible k-mers is bounded by the sequence length $(n-k+1)$. This also makes it computationally expensive to analyze the counts of k-mers for larger values of k.

1.6 Markov chains

There is a close correspondence between the count of k-mers and the probabilistic models called Markov chains. The Markov chains are an extension of an independent and identically distributed (i.i.d.) model, which captures the probability of observing either of the A, C, G, and T nucleotides in a sequence. This corresponds exactly to the relative frequency of the 1-mers ($f_{\mathtt{A}}$, $f_{\mathtt{C}}$, $f_{\mathtt{G}}$, $f_{\mathtt{T}}$). In these types of models, the probability of observing each nucleotide is assumed to be independent of all other observed characters.

The higher-order Markov chain expands on the i.i.d. model by modeling the probability of observing a character c after a k-mer w. With $k = 1$, w is an individual character (A, C, G, or T), and the model is a first-order Markov chain. Note that 'Markov chain' without further qualifications refers to a first-order Markov chain. In higher-order Markov chains, the k-mers w, where $k > 1$, each correspond to a state of the model, and the model contains the probabilities of observing each nucleotide after each k-mer. Both c and w are composed of characters from an alphabet Σ, which in the context of DNA is typically $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, with $c \in \Sigma$ and $w \in \Sigma^*$. We denote the probability as $p(c|w)$, which can be estimated based on the frequency f_w of w and the combined k-mer wc, as $\hat{p}(c|w) = f_{wc}/f_w$.

Specifically, each set of k-mers corresponds to a higher-order Markov chain with an order of $(k-1)$. Each probability of the higher-order Markov chain can be estimated from the k-mer counts as $\hat{p}(c|w) = f_{wc}/\sum_{\sigma\in\Sigma} f_{w\sigma}$. See Figure 1.8 for an example.

1.6.1 Variable-length Markov chain

The variable-length Markov chains (VLMCs) are an extended model of the higher-order Markov chains [118, 119]. Where the higher-order Markov chains has a fixed size of the k-mers, in the VLMCs, the size of the k-mer in each state is allowed to vary. This makes the VLMCs more flexible, as they can include shorter k-mers when, e.g., a k-mer is infrequent and thus the probabilities cannot be estimated well, but can have longer k-mers where that improves the specificity of the model. Consider, for example, the sequencing errors that are introduced in sequencing experiments. On

Higher-order Markov chain				Variable-order Markov chain			
k-mer	c	Count	$\hat{p}(c\|k\text{-mer})$	k-mer	c	Count	$\hat{p}(c\|k\text{-mer})$
AAAA		190		AAAA		190	
AAAA	A	70	$70/190 \approx 0.38$	AAAA	A	70	$70/190 \approx 0.38$
AAAA	C	20	$20/190 \approx 0.11$	AAAA	C	20	$20/190 \approx 0.11$
AAAA	G	40	$40/190 \approx 0.21$	AAAA	G	40	$40/190 \approx 0.21$
AAAA	T	60	$60/190 \approx 0.32$	AAAA	T	60	$60/190 \approx 0.32$
AAAC		200		AAAC		200	
AAAC	A	30	$30/200 = 0.15$	AAAC	A	30	$30/200 = 0.15$
		...				...	
CGAG		10		GAG		101	
CGAG	A	3	$3/10 = 0.3$	GAG	A	33	$33/101 \approx 0.33$
CGAG	C	3	$3/10 = 0.3$	GAG	C	23	$23/101 \approx 0.23$
CGAG	G	0	$0/10 = 0.0$	GAG	G	14	$14/101 \approx 0.14$
CGAG	T	4	$4/10 = 0.4$	GAG	T	40	$40/101 \approx 0.40$
		...				...	

Figure 1.8 Relationship between k-mers and the probabilities in Markov chains. Estimating the probabilities of A, C, G, T following the k-mer CGAG is complicated by the low frequency of the k-mer, but with variable-length Markov chains, it is possible to exclude the k-mer and instead use the probabilities following GAG. The small counts of k-mers makes the estimated probabilities vulnerable to potential errors in the sequence, such as those from sequencing errors.

infrequent k-mers, such errors will have a large impact on the estimated probabilities. Therefore, by excluding infrequent k-mers, the models can be more robust.

There are many ways to select which k-mers are included in a VLMC [120–126]. However, the standard method is based on the Kullback-Leibler divergence [126]. In this method, each state with k-mer cw is compared to the more general state with $(k-1)$-mer w, which is one shorter and typically referred to as the parent of the cw state. The divergence is computed as

$$\Delta_{cw} = f_{cw} \sum_{\sigma \in \Sigma} \hat{p}(\sigma|cw) \log \left(\frac{\hat{p}(\sigma|cw)}{\hat{p}(\sigma|w)} \right). \tag{1.1}$$

States with k-mers cw that have a $\Delta_{cw} < K$ are removed from the VLMC. The value of K is referred to as the threshold and is used to determine the size of the model, with smaller values giving larger VLMCs.

1.6.2 Advantages of variable-length Markov chains

Compared to the k-mer or higher-order Markov chain approaches, the VLMC includes a principled way to constrain the model's size. For k-mers and higher-order Markov chains, a fixed k value has to be set, which always gives 4^k parameters in the model. As k increases, the number of parameters increases exponentially, and many k-mers will not even occur once. Since the VLMCs do not have to include all k-mers

for a fixed k, the number of parameters can instead vary depending on which can be estimated with sufficient support. For example, suppose a k-mer is infrequent in a given sequence. In that case, the VLMC can instead include the $(k-1)$-mers where the corresponding probabilities can be accurately estimated (e.g., Figure 1.8).

1.6.3 Applications of variable-length Markov chains

Variable-length Markov chains have previously been applied to study various biological questions. Example applications include the detection of horizontal gene transfer in bacterial genomes [127], tracing the origin of plasmids [102], and predictions of indel flanking regions in proteins [128]. More recently, VLMCs were applied in conjunction with k-mers to better approximate the background adjustment of the d_2^* measure when clustering transcriptomes [129].

To perform these analyses, a VLMC is computed for each genome in a collection. Then, the likelihood of each query sequence is computed for all VLMCs. The VLMC that gave the highest likelihood is identified, indicating which genome the query sequence is the most similar to.

1.6.4 Other variable *k*-mer approaches

Variable-length Markov chains are not the only alignment-free method that incorporates k-mers of different lengths. Other methods that account for k-mers of variable sizes include `EP-SIM` [130], which targets short k-mers, and `PopPUNK` [131], which targets longer k-mers. `EP-SIM` [130] achieves this by using Entropic Profiles, which combines information about the distribution of k-mers up to a fixed length in the entire genome with the distribution of k-mers at each position in the genome. `PopPUNK` [131] instead uses every other k-mer size between two ranges k_{min} and k_{max} (default $= 29$), and computes how many such k-mers are shared between two genomes. In addition, an early definition of D_2 suggested that it might be useful to include the count of k-mers of multiple lengths, where $l \leq k \leq u$ for some l and u [111].

1.6.5 Summary of Markov chains

In summary, higher-order Markov chains, and in particular variable-length Markov chains, are methods that capture the specifics of sequences without relying on sequence homology. They can be applied to classify sequences of unknown origin or to study evolutionary relationships between sequences. Variable-length Markov chains improve on higher-order Markov chains and k-mers by allowing for flexibility in which k-mers are included in the models.

1.7 Genomic signatures

As previously discussed, certain organisms have been shown to have a specific preferred usage of, for example, dinucleotides [74, 75] and codons [83]. In addition, early

work on k-mers showed that it is possible to distinguish between coding and non-coding regions by using the counts of 3-mers [132]. Although genomic signatures were originally defined only on dinucleotides, we extend this definition to include k-mers of variable lengths. By analyzing longer k-mers, such as 6- or 9-mers, it is likely possible to observe more specific genomic signatures than with dinucleotides. The specific patterns of k-mers will be biased by, for example, preferences for specific dinucleotides, codons, or codon pairs. In addition to these biases, they may also be influenced by specific binding sites or transcription factors, or other mechanisms acting on their genomes.

Many virus genomes have a nucleotide composition that is similar to the composition of their hosts [133]. However, this is not the case for all viruses, as exemplified by the two alphaherpesviruses HSV-2 and VZV, which both infect humans but have a large difference in GC content: 70 %, and 46 %, respectively. In addition, the dinucleotide content of some plasmids is similar to their hosts, while plasmids with broad host ranges are often not similar to any of their hosts [75].

Furthermore, it has been shown that it is possible to alter the fitness of viruses by altering their codon usage, for example, to increase gene expression of HPV 16 [57] and HPV 11 [134] and to attenuate the ΦX174 phage [135]. This highlights that specific codon usage is important for the expression of viral genes. Despite this, it was recently shown that most eukaryotic viruses' genes are not adapted to be similar to the host's average codon usage [133].

Codon pairs have also been illustrated to be similar between the polio virus and the human genome, and can be used to attenuate the virus [88]. However, the frequency of codon pairs has also been illustrated to strongly correlate with the frequency of dinucleotides [136]. In addition, for several viruses that were attenuated by altering their codon pairs, the cause of the attenuation is thought to be an increase of the specific dinucleotides that are targeted by the host immune system [137–139]. However, altering the codon pairs has also been illustrated to have an attenuating effect on influenza A viruses, while specific changes in dinucleotides did not attenuate the viruses [87]. Furthermore, both codon pairs and dinucleotides have been shown to only marginally be similar to the host of viruses, which implies that there may not be a large degree of adaptation of either [136, 140].

As previously discussed, the analysis of biological properties of genomes, such as the GC content or dinucleotides, falls into the domain of alignment-free methods. For instance, the counts of 1-mers (f_A, f_C, f_G, f_T) corresponds to the GC content of the genomes ($(f_C + f_G)/(f_C + f_G + f_A + f_T)$). Likewise, comparing 2-mers (e.g., f_{CG}) is equal to comparing dinucleotides of the genomes. Parts of all 3-mers in a genome include the codons of the genome. Note that in the typical method, k-mers disregards genes' reading frame and counts all possible 3-mers.

Recently, the 5-mer counts of archaea genomes were found to correlate both with the phylogeny of the genomes and the salinity and temperature of the ecological niche of the archaea [141]. They also showed that many archaeal viruses cluster primarily according to their taxonomy rather than with their hosts. This result is in contrast to other studies on bacteriophages' host similarity, where many viruses that infect bacteria were shown to be similar to their hosts [113, 142].

In addition, the counts of 3-mers of ancient integrated retroviruses in the human genome were recently analyzed [143]. They found that most integrated virus genomes still had a similar 3-mer composition to extant viruses and were not, over time, significantly altered to have human-like 3-mer counts. Furthermore, by using 3-mer counts, they identified several previously undiscovered inserted virus-like elements in the human genome.

In summary, all of these findings support that there are preferences for various biological properties in the genomes of viruses, with implications for their fitness. These specific preferences for a certain set of k-mers in the genome of a species are referred to as the genomic signature of that organism. Note that the k-mer preferences that we discuss here are those that are subject to selective pressure, which causes a certain set of k-mers to be present throughout a genome.

Furthermore, it might be possible to illustrate more specific genomic signatures with the help of k-mers instead of dinucleotides. To address some of the issues with k-mers, we use the variable-length Markov chain for the analysis of genomic signatures, as previously suggested [127].

Chapter 2 Aim

To our knowledge, no comprehensive analysis of the genomic signatures in viruses exists. Specifically, it is not known if there are specific preferences of k-mer counts throughout the genomes of viruses, as has been previously shown for eukaryotes and prokaryotes with dinucleotides. Thus, the aim of this thesis was to study the evolution and host-adaptation of virus genomes, with a focus on genomic signatures. The first part of this aim was to develop methods that make analysis of genomic signatures easy to perform and computationally efficient (paper I, II). With these improved algorithms, the second part of the aim was to thoroughly investigate how common and specific genomic signatures are in viruses (paper III). The third part was to establish to which extent the genomic signatures of viruses are adapted to the genomic signatures of their hosts (paper IV). The final part of the aim was to demonstrate how genomic signatures can be used to answer important evolutionary questions by applying them to detect recombination (paper V).

Chapter 3 Methods

3.1 Variable-length Markov chains

The construction of VLMCs is a time and memory-intensive task. There are several algorithms for the construction of VLMCs from the early 2000s [122–124], but none are able to handle modern datasets or sequence sizes. An improved but no longer available method from 2008 achieved increased memory and speed efficiency [144, 145]. Thus, there is a need for new algorithms for VLMCs that can handle modern datasets.

One traditional application of the VLMC is to predict the likelihood of a sequence given a specific VLMC. This can be applied, for example, to classify unknown sequences in metagenomic studies. For a given sequence S and VLMC λ with the maximum length of the k-mers as k, the likelihood is computed as

$$L(S|M) = \prod_{i=0}^{|S|} p(S_i|S_{i-k} \dots S_{i-1}, \lambda). \tag{3.1}$$

Here $p(S_i|S_{i-k}...S_{i-1}, \lambda)$ refers to the likelihood of the ith nucleotide in S after observing the k previous nucleotides. For computational purposes, this is typically log-transformed and then negated to give positive values,

$$-\sum_{i=0}^{|S|} \log \hat{p}(S_i|S_{i-k} \dots S_{i-1}, \lambda). \tag{3.2}$$

This gives a measurement called the negative log-likelihood, which is sometimes referred to as the score of a sequence.

Since the likelihood and negative log-likelihood of a sequence depends on the length of the sequence, with a longer sequence giving a larger value, we usually use the normalized likelihood. This is computed by calculating the negative log-likelihood of a sequence and dividing the result by the length of the sequence. This is then transformed into a likelihood by computing the exponential of the result.

A further consideration is the correlation between GC content and the likelihood under a VLMC. This correlation indicates a tendency to assign a large likelihood to unrelated sequences due to a similarity in GC content. Correcting for the GC content is

straightforward, as the GC content of a sequence can be found for any VLMC. Specifically, the GC-adjusted negative log-likelihood of a sequence S can be computed as

$$\sum_{i=0}^{|S|} \log \frac{\hat{p}(S_i|S_{i-k}\dots S_{i-1},\lambda)}{\sqrt{\hat{p}(S_i|\lambda)}}, \tag{3.3}$$

where k is the maximal length of a k-mer in the model. This adjusted version no longer recognizes sequences with similar GC content as similar unless they also share similarities in longer k-mers.

We additionally apply sliding window analyses to study the genomic signatures in viruses. The sliding window approach refers to the computation of the negative log-likelihood of short, overlapping parts of the sequence (windows). Typically, this is visualized as a window that is moved (or slid) over the sequence. We usually show the results as the normalized likelihood of the sequence of each VLMC for each window.

3.1.1 Parameter selection of VLMCs

Selecting which parameters to use when training VLMCs depends on the specific task. The specific parameter selection method can vary. In the following, we offer some guidance.

There are three parameters that can be tuned when constructing VLMC: the maximum length of an included k-mer, the minimum count to include a k-mer and the Kullback-Leiber threshold. The most important parameter is the threshold of the Kullback-Leibler-based comparison between each parent and child state Δ_{cw} (Equation (1.1)). In VLMCs, the parent state is the $k-1$ suffix of the child k-mer. For example, the child state `ACG` would have a parent state that corresponds to `CG`. The value of Δ_{cw} correlates roughly with how different the probabilities of the two states are and with the count of the corresponding k-mer (Figure 3.1).

One approach to selecting parameters of a model is the Bayesian information criterion (BIC) [146]. The BIC of a specific model is defined as

$$\text{BIC} = \text{card}(\lambda)\ln|S| - 2\ln L(S|\lambda), \tag{3.4}$$

where S is the sequence that the VLMC λ is trained on, $|S|$ is the length of the sequence, and $\text{card}(\lambda)$ is the size or number of parameters of the model. We compute $\text{card}(\lambda)$ from the number of (terminal) states in the model λ multiplied by the number of free parameters per state (for DNA, there are three free parameters). By computing the BIC on a large set of short and medium-sized sequences, we found that the optimal value given by the BIC of Δ_{cw} increases linearly with sequence size (Figure 3.2).

From the same analysis, we also find that the optimal maximal length of k-mers included in the models also grows with the sequence length (Figure 3.2). For most cases, the optimal maximal depth is below 10, even for sequences 10 million nt long.

Furthermore, the size of the VLMCs grows linearly with an increase in sequence length (Figure 3.3). Here, the size of a VLMC is defined as the sum of the length of all included k-mers. This illustrates the flexibility of the VLMCs, as they are adapted to the length of the sequence.

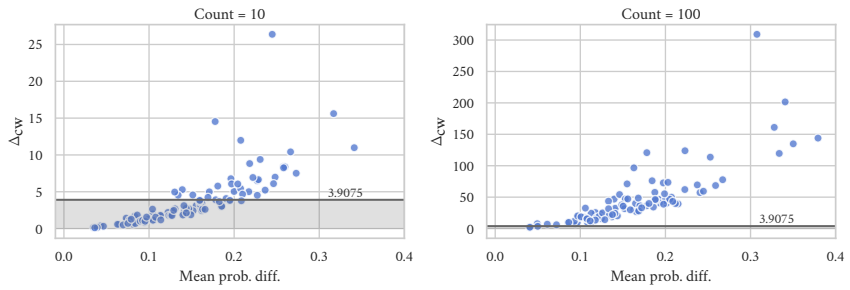

Figure 3.1 Illustration of Δ_{cw} as a function of the difference in probabilities. Each dot corresponds to a state with a count of either 10 (on the left) or 100 (on the right). The Mean prob. diff. refers to the average absolut difference in the probabilities of the child and parent state. With $K = 3.9075$, all states in the shaded area will be removed.

For other tasks, such as the construction of phylogenetic trees to analyze the evolutionary relationships between sequences, setting Δ_{cw} close to 0 gave trees that better corresponded to the trees of similar methods. This highlights that it is important to select appropriate parameters based on the task.

3.1.2 Graphical representation of VLMCs

To enable easy visual inspection of VLMCs, we have also developed a graphical representation. The representation is based on the probabilities of the models, and the structure of the representation shows the included k-mers in the VLMCs. Thus, the shape of the representation allows for visual inspection of the differences in genomic content between sequences.

By using the relative frequencies present in the VLMC, we construct a representation of the k-mers the model represents. In the most general case, the VLMCs contain only the probabilities of the single nucleotides A, C, G, and T, and the visualization thus contains the probability of each nucleotide (Figure 3.4A). As we increase the complexity of the model, the VLMCs capture the frequencies of dinucleotides (Figure 3.4B). In the representation, the dinucleotide AC follows C, as AC is an extension of the C context. The size of the dinucleotide fraction in the visualization corresponds to the likelihood of observing the dinucleotide in the sequence. Including 3-mers extends the visualization in an analogous manner (Figure 3.4C). The size of 3-mer in the visualization corresponds to the likelihood of observing the 3-mer in the sequence and is placed after its suffix to illustrate the growing contexts in the VLMCs. During the VLMC construction, however, some k-mers are pruned, and thus the corresponding spaces in the visualization are left blank, as can be seen for 3-mers in Figure 3.4C.

We illustrate an application of the representation by computing the VLMCs of the *Human alphaherpesvirus 1* (HHV-1) and the *Human alphaherpesvirus 3* (HHV-3) (Figure 3.4D, E). By looking at the illustration of HHV-1, it is clear that it has a GC-rich genome, as the fractions of G and C dominate. Further, the CC, CG, GC, and GG dinucleotides occupy a large portion of the dinucleotide space. The long k-mers in the

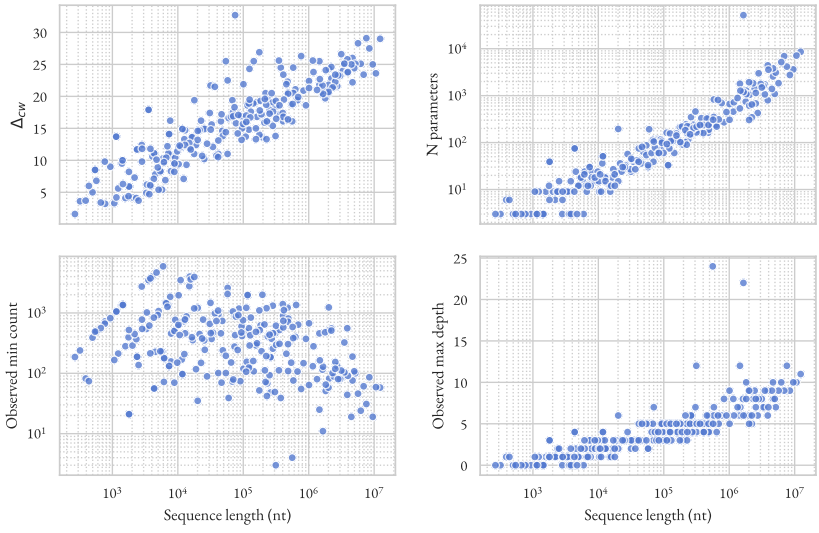

Figure 3.2 Illustration of the optimal parameters of the VLMC creation. The optimality is defined by the BIC. The observed min count and max depth are the values of the state in the VLMC with the smallest count and longest k-mer respectively.

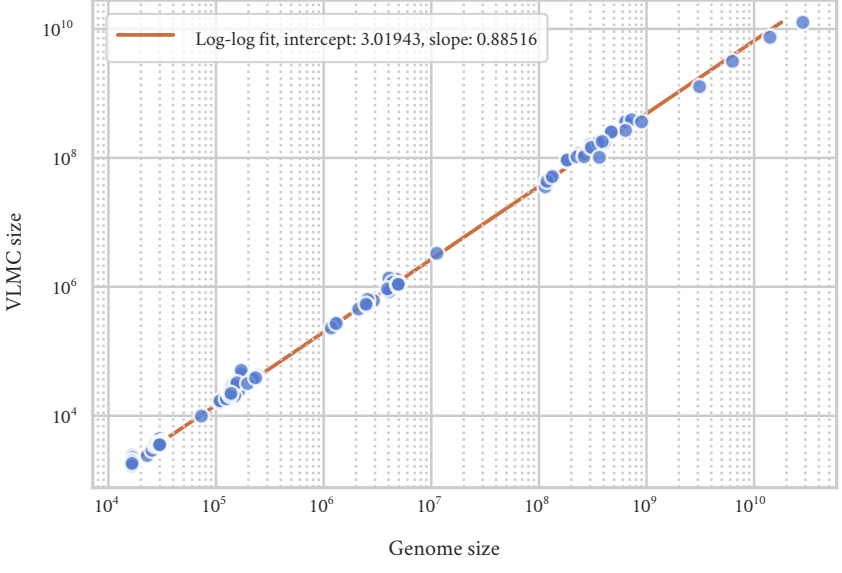

Figure 3.3 The size of VLMCs correlates with the genome size. With fixed parameters settings of $\Delta_{cw} = 3.9075$, the size of VLMCs grow linearly with the genome size.

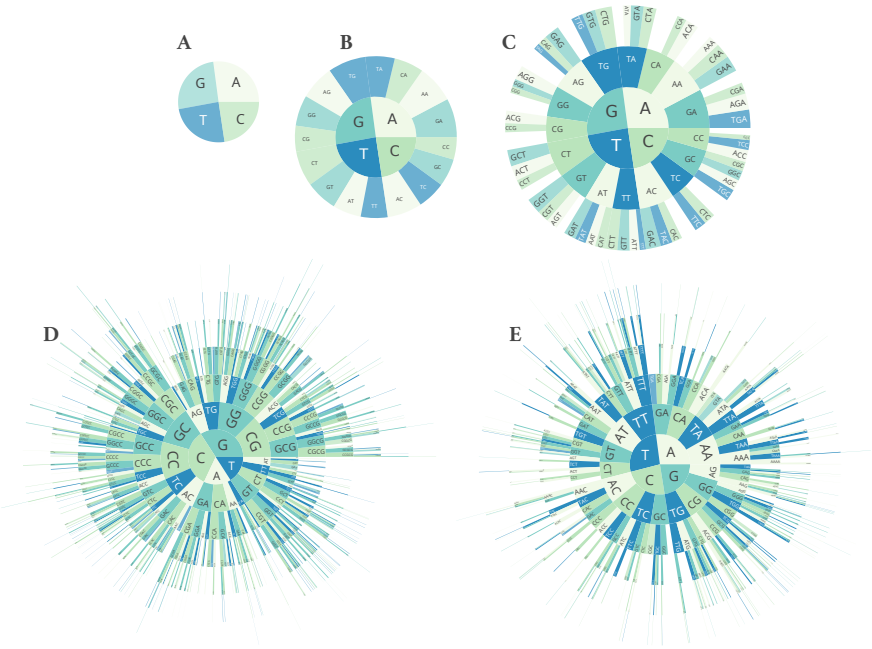

Figure 3.4 Graphical representation of VLMCs. Each area in the visualization contains the conditional probability of observing that context in the VLMC. A, B, and C contains an increasingly more complex VLMC computed on the sequence NC_001416.1, while D and E are computed on HHV-1 and HHV-3, respectively (see text for details). The different shapes of the models allow for visual inspection of the differences in their respective genomes.

model also correspond to branches of GC-rich k-mers, with an occasional `T` or `A`. In contrast, the illustration of the HHV-3 genome shows a relatively balanced nucleotide content, with a slight bias towards `A` and `T`. A few long branches start at repeats such as `AAAA` or `ATAT`. The visualizations show that these two viruses, although both Herpesviruses infect humans, are different.

3.2 Datasets

The datasets of each paper are described in the respective paper. As a summary, to benchmark the speedup and memory usage of the algorithms in paper I and II, we used some of the largest known genomes, with representatives from most domains (Table 3.1). In paper III, we used a dataset of eukaryotic viruses. The included sequences were selected based on the current ICTV classification of viruses (Table 3.2). In paper IV, we additionally included prokaryotic viruses as well as a large collection of host genomes (Table 3.3).

Organism	Accession	Length (Mbp)	GC %
Pandoravirus salinus	NC_022098.1	2.474	61.72
Sorangium cellulosum	GCF_004135735.1	11.261	72.58
Drosophila melanogaster	GCA_004798055.1	133.404	42.12
Oryza sativa	GCA_001623365.2	387.424	43.61
Symbiodinium kawagutii	GCA_009767595.1	935.067	45.54
Homo sapiens	GCA_000001405.28	3099.706	41.04
Palaemon carinicauda	GCA_004011675.1	6699.724	37.37
Pinus taeda	GCA_000404065.3	22 103.636	37.45
Ambystoma mexicanum	GCA_002915635.2	32 396.370	44.97
Neoceratodus forsteri	GCA_016271365.1	34 557.648	44.13

Table 3.1 Genomes used to benchmark the algorithms in papers I and II. The Accession is the NCBI accession id of the corresponding assembly or sequence for the organism.

Baltimore	Number of sequences	Number of species	Number of families
dsDNA	818	528	18
dsDNA-RT	89	89	2
dsRNA	717	135	10
ssDNA	895	711	8
(+)ssRNA	1107	926	44
(-)ssRNA	567	306	21
ssRNA-RT	53	53	1

Table 3.2 Summary of dataset used in paper III. The number of sequences are sometimes much larger than the number of species due to the segmented viral genomes.

Group	Number of sequences	Number of species	Number of families
Archaea	445	140	29
Bacteria	5491	1483	327
Fungi	1803	167	54
Insecta	7873	399	99
Metazoa	12 821	530	293
Viridiplantae	5528	391	98
Viruses	12 272	9338	179

Table 3.3 Summary of dataset used in paper IV. The number of sequences include both the segments of segmented viruses and the chromosomes of prokaryotes and eukaryotes.

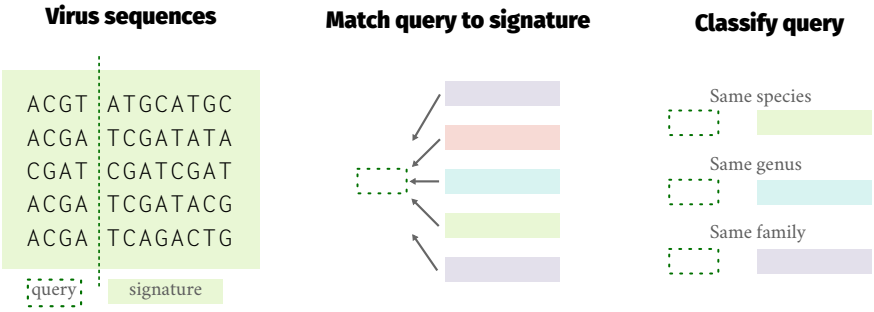

Figure 3.5 The method used to determine if a virus has a specific genomic signature. Each virus genome is split into two parts, one containing 30 % of the genome (the query) and the other 70 % of the genome (the signature). For each query, we find the signature that is the most similar to it based on the likelihood of the sequence.

3.3 Method of detecting genomic signatures

To analyze to what degree viruses have specific genomic signatures, we studied if each virus sequence had a specific pattern of k-mers throughout its sequence. To do this, we split each virus sequence into two parts, the first 30 %, termed query, and the last 70 %, termed signature (Figure 3.5). By comparing the two parts with VLMCs computed on the 70 % and likelihoods computed on the 30 %, this illustrated in which cases the patterns of k-mers present in both parts were similar. A more thorough test could instead test arbitrary parts of the genome against each other. We found that this more thorough and time-consuming test gave similar results as the simple test. To not bias these results by repeat regions, simple repeat regions were removed prior to the analysis.

We designed a statistical test to verify that the observed results would not be observed simply due to the number of viruses in each genus or family. This test compared the number of observed viruses with specific signatures to the number of viruses with specific signatures that would be expected at random. The distributions were compared with a one-sided t-test with an alpha of 0.05 and adjusted with the Bonferroni method.

3.4 Genomic signatures of viruses compared to their hosts

We further analyzed how similar viruses' signatures are to their hosts' signatures in paper IV. To do this, we designed a statistical test based on dissimilarities between viruses and their hosts compared to those between viruses and all other members of the host's domain. We expanded this by also comparing to members of the virus' host's family, order, and phylum (Figure 3.6).

The second test of host adaptation of viruses' signatures compared viruses with

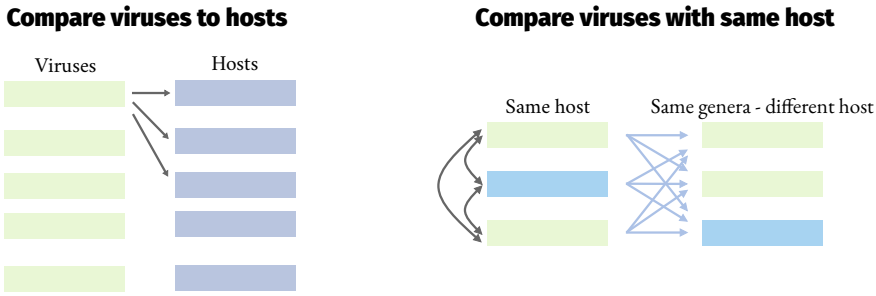

Figure 3.6 The methods of analyzing host adaptation in paper 4. Dissimilarities between viruses' signatures and hosts' signatures were compared to determine if viruses had adapted their signatures based on similar selection pressures as those acting on the hosts' signatures. The dissimilarities between signatures of viruses with the same host were compared to the dissimilarities of related viruses with other hosts.

the same host. Here, we compared the distribution of dissimilarities of viruses with the same host to the distribution of dissimilarities between those viruses and all other members of the same genera with different hosts.

Both analyses were implemented with Mann-Whitney U tests [147] as the dissimilarity distributions were not normally distributed. The p-values corrected for multiple hypotheses with the Benjamini/Hochberg method [148].

3.5 Sequencing

In paper V, we developed a method to detect recombination events from sequencing data. We provide the exact details in the paper. In the following, we provide the reasoning behind the choice of sequencing. We sequenced both virus strains grown in laboratory conditions and a sample from a patient. To do the sequencing, we opted for nanopore sequencing. Primarily, recombination events are easier to classify with long reads, which the nanopore sequencer can provide. Specifically, it is easier to detect a recombination event if there is a sufficient amount of sequence on both sides of the recombination breakpoint. Therefore, the longer the reads, the easier to detect recombination. We selected nanopore over other long-read sequencing methods because we could easily run the methods ourselves.

Chapter 4 Results

4.1 Algorithms for variable-length Markov chains

The aim of paper I was to improve the algorithms for the construction of variable-length Markov chains. Previous algorithms for this task were slow and could not handle even moderately sized bacterial genomes due to the large amounts of memory required. This was identified as a significant issue as we were interested in comparing the genomic signatures of viruses to their hosts (see paper IV).

To speed up the VLMC construction, we re-implemented a faster algorithm that was no longer available [144]. This algorithm is based on a data structure called lazy suffix trees [149]. We improved this algorithm by designing a parallelized construction scheme. In addition, some further optimizations were realised with hash maps.

The new method is significantly faster and requires up to 1000 times less memory than the original algorithm. Compared to `VOMM` [150], the previous state-of-the-art, the new algorithm was up to 100 times faster for common parameter choices of VLMCs (Table 4.1). For example, computing a VLMC on the Human genome with `VOMM` takes 3.3 hours, and with our algorithm, 2.2 minutes. However, the memory usage was slightly larger at $\approx$ 19 bytes per character, while `VOMM` uses $\approx$ 12 bytes per character. For example, computing a VLMC on the Human genome takes $\approx$ 63GB of RAM, while `VOMM` uses $\approx$ 39GB of RAM. On larger genomes, the parallelized algorithm achieves a speedup of close to 9 on 32 cores. As 92 % of the algorithm can run in parallel, this is close to the best possible speedup according to Amdahl's law [151].

We also extensively tested which parameters to use for the training of VLMCs based on the BIC parameter selection method. We found that the BIC gave an optimal depth of the model and Δ_{cw} threshold that increases with sequence size, but the optimal min count does not, and for the longer genomes was often above 100. This model selection is further explored here on a larger set of sequences in Section 3.1.1.

In conclusion, our improved algorithm enables the computation of VLMCs on all prokaryotic genomes and most eukaryotes, except for the current two largest sequenced genomes. With parameters determined by the BIC, the method is up to 100 times faster than the `VOMM`. In addition, it was the first parallel algorithm for the construction of VLMCs. We also implemented a parallelized computation of the negative log-likelihood (see Equation (3.2)) and the computation of the negative log-likelihood of adjacent subsequences of a sequence (sliding window).

4.2 External-memory construction

The relatively high memory usage of the previous algorithm prohibited its use on larger genomes. While such genomes could have been excluded here, the time to compute VLMCs on moderately sized genomes for parameters larger than the ones suggested by the BIC was still significant. For example, increasing the maximum value of k in the model from 8 to 15 resulted in 14 times slower construction. Therefore, the aim of paper II was to develop an algorithm that could compute VLMCs for larger values of k and longer genomes.

The algorithm is based on k-mer counts, which can be computed for genomes and sequencing data sets with several efficient methods [152–154]. From these k-mer counts, all of the l-mer counts for $l \leq k$ are computed iteratively. And, as illustrated in fig. 1.8, from these l-mers, it is possible to estimate a variable-length Markov chain. The l-mers are then sorted to enable the computation of Δ_{cw}, so that each parent w (e.g., `CGT`) is proceeded by all its child cw states (e.g., `ACGT`).

The algorithm is up to 70 times faster than our previous algorithm. In addition, it is more memory efficient than our previous algorithm and can also optionally run in external memory, where data is stored on disk instead of in RAM. While running in external memory is approximately two times slower than the RAM implementation, it enables the construction of even the largest sequenced genomes.

These computational improvements also allow VLMCs to be computed directly on sequencing read data. We illustrated that VLMCs computed on sequencing reads are similar to VLMCs computed on reference genomes. This enables applications where VLMCs are computed directly on sequencing data, bypassing assembly.

We also developed two dissimilarity measurements for VLMCs. We call them d_v and d_v^*, which compare two VLMCs faster than the negative log-likelihood. The methods are similar to the k-mer method d_2^* [112]. We used d_v^* to construct phylogenetic trees from sequencing reads, with slightly better results than other methods, particularly on higher sequencing error rates.

In conclusion, this algorithm enables the computation of VLMCs on genomes of all sizes. These improvements further enable VLMCs and genomic signatures to be computed directly on sequencing data, allowing computations even on most laptops.

4.3 Genomic signatures in viruses

Powered by these improved algorithms, the aim of paper III was to map to which extent viruses have specific genomic signatures. It has previously been illustrated that many prokaryotes and eukaryotes have specific dinucleotide content [74, 75]. In particular, the dinucleotide contents of many of these organisms are preserved in much of the genomes. This might indicate a selective pressure acting on the genomes of these organisms, where the specific dinucleotide content confers some evolutionary advantage. Since viruses often rely on their host's genetic and translational machineries for their replication and gene expression, viruses might be subject to similar selective pressure acting on their genomes as their hosts. However, it is not known if viruses

Organism	Size (Mnt)	Method		
		VOMM	Paper I	Paper II
Pandoravirus salinus	2.474	5.4	0.11	0.40
Sorangium cellulosum	11.261	24.9	0.30	0.48
Drosophila melanogaster	133.404	400	3.25	0.85
Oryza sativa	387.424	1251	13.6	6
Symbiodinium kawagutii	935.067	3116	27	15
Homo sapiens	3099.706	12 039	131	39
Palaemon carinicauda	6699.724	22 209	368	73
Pinus taeda	22 103.636	75 289	8895	130
Ambystoma mexicanum	32 396.370			*462
Neoceratodus forsteri	34 557.648			*531

Table 4.1 The time in seconds to construct a VLMC with the parameters given by the BIC. VOMM refers to the algorithm from [150], which constitutes the state-of-the-art, and Paper I refers to the HashMap version presented there and II to the in-memory version of the algorithm unless otherwise stated. * refers to the use of the external-memory version. Empty cells correspond to the corresponding algorithm using more than 360 GB of RAM or more than 24 hours.

have specific genomic signatures or to what extent such signatures are similar to the signatures of their hosts.

To analyze genomic signatures in viruses, we compared the genomic signatures of the last 70 % of each genome to the first 30 % of genomes. This gives a method of investigating if there is a specific bias in nucleotide content, dinucleotides, codons, and codon pairs throughout the genome. Each virus was assigned as having either a species-, genus-, or family-specific signature, depending on which 30 % part was the most similar to the genomic signature computed on the last 70 % of the genomes. See Figure 3.5 for an illustration.

We based the genomic signatures on variable-length Markov chains (VLMCs) with a maximal depth of 6, which captures up to 7-mers in the genomes. Choosing an appropriate size of the VLMCs is crucial to accurately capture the genomic signatures, which is done by picking a value of Δ_{cw} (Equation (1.1)). Small values of Δ_{cw} make the models capture the sequence used to train the model well (the last 70 %), but the model is too specific, which leads to a low likelihood of the first 30 % (Figure 4.1). With larger values of Δ_{cw}, the VLMCs become too general and cannot distinguish between the sequence and unrelated sequences. We found that setting Δ_{cw} to 3.9075, as suggested by the documentation of an earlier VLMC algorithm [127] (corresponds to half of the chi-squared distribution with 3 degrees of freedom at a p-value of 0.05) is a reasonable tradeoff between these factors. The average (per family) length of k-mers in the VLMCs ranged from 1 (which captures dinucleotides) to close to 6 (captures 7-mers).

By thoroughly analyzing genomic signatures in viruses, our method revealed that most viruses have genomic signatures that are at least family-specific. In addition, most viruses with genomes longer than 50 000 nt had species-specific signatures. In

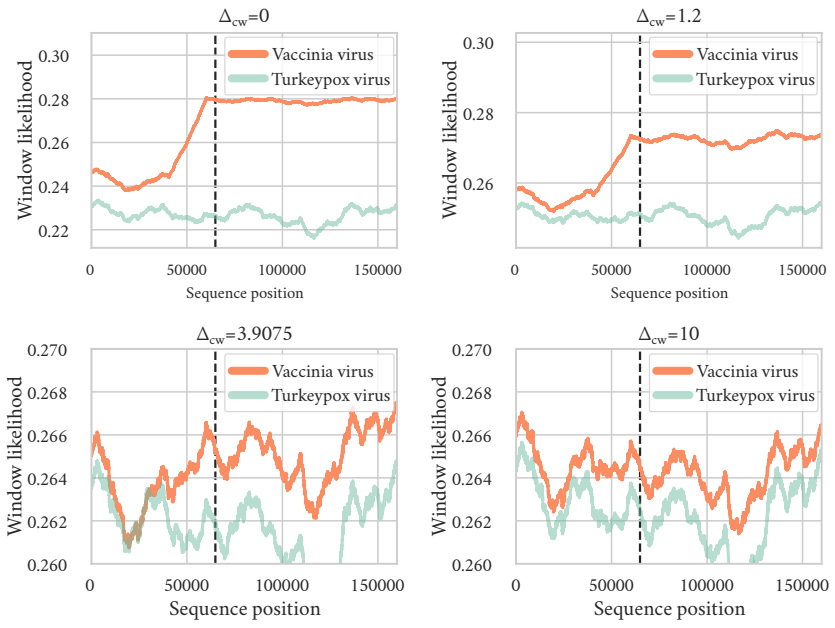

Figure 4.1 Illustration of parameter selection for the detection of genomic signatures. The sequence to the right of the dotted line is used to train the VLMC. For small values of Δ_{cw} (0, 1.2), there is a marked difference between the average window likelihood of the sequence used to train the VLMC and the first 30 % of the sequence. For larger values of Δ_{cw} (10), the specificity of the model decreases, and the likelihood of unrelated sequences approaches the likelihood of the training sequence.

contrast, many short viruses had either genus- or family-specific signatures or no discernable signatures (Figure 4.2). Likewise, most viruses had specific signatures in the dsDNA group (which contains the longest viruses), while the other Baltimore classification groups had no specific trend regarding DNA or RNA viruses. In all cases, the results were statistically verified.

We further developed a sliding window protocol to analyze the presence of signatures throughout each genome. This sliding window protocol compares windows of lengths from 50 nt to 10 000 nt of each genome to the genomic signatures of all other genomes. Each window is also compared to the signature of the source genome, but the window is removed from the genome prior to computing the signature. By analyzing these windows of each genome, we found that for many viruses, the signatures are present in most or all of the genome. In some cases, certain regions had family-specific signatures, while the rest of the genome was species-specific, which might indicate highly conserved regions. There were also some areas where the signatures could not be detected, most notably in regions with many repeats.

In addition, we analyzed how similar the signatures of each viral family were by constructing a neighbor-joining tree. There, we found that many of the viruses with

genus- or family-specific signatures also had similar signatures to most of the members of their genus or family. This is expected as if they had not had similar signatures, they would not be genus- or family-specific. However, for many of the species-specific genomic signatures, the signatures were often not similar to all signatures from the same genus or family. Instead, there were many smaller groups of similar signatures, often from the same family but not necessarily from the same genus.

We also investigated if the genomic signatures of viruses were very similar to the signatures of their hosts. To do this, we analyzed if each virus' signature was the most similar to the signature of one of its hosts or a closely related host. Our results show that only a small set of viruses were similar to their host.

In conclusion, most viruses have specific genomic signatures. In addition, many closely related species have similar signatures, but many viruses also have distinct signatures. In addition, the signatures are present in the entire genomes, which indicates that a selective pressure acting on the genomes of viruses gives rise to the genomic signature. However, only a few viruses had signatures similar to their host's signatures, which shows there are different preferences and likely a different selective pressure acting on the signatures of hosts and viruses.

4.4 Virus-host similarities in genomic signatures

In paper III, we found that most viruses did not have genomic signatures similar to their host's signature. However, the assumed adaptation of viruses' genomic signatures to their host's signatures could be more subtle and compete with other selective pressure influencing the signatures. Therefore, the similarity between the genomic signatures of viruses and hosts might be more subtle than the previous method could detect. Thus, the aim of paper IV was to study whether each virus' signature is subtly similar to its host(s) signature. We designed a statistical test of similarity with the null hypothesis that each virus' signature is as similar to its host(s) as it is to all organisms in the same phylum or kingdom as the host. To separate GC content from the rest of the signatures, we used the d_v^* dissimilarity developed in paper II. As a comparison, we included the organisms' nucleotide, dinucleotide, codon, and codon-pair usage.

We found that most viruses' signatures are not more similar to their host's signatures than expected, other than in GC content. There were a few exceptions, notably, all of the endogenous viruses *Polydnaviridae*, 11 % of the *Flaviviridae*, 6 % of the *Coronaviridae*, 6 % of the *Adenoviridae*, and 6 % of the *Geminiviridae*. On average, more viruses with insect hosts were more similar to their hosts than viruses with other hosts, including bacteria and archaea.

In contrast, the nucleotide content of most viruses was more similar to their hosts than expected. In particular, close to 90 % of Archaea and Bacterial viruses had a similar GC content to either their hosts or a closely related host. For other viruses, close to 60 % of viruses were similar to a related host in GC content. For dinucleotides, close to 40 % of insect and non-insect metazoa viruses were similar to their hosts. For codons and codon pairs $\approx$ 20 % to 30 % of viruses were similar to a related host.

We further analyzed the similarity of signatures within a few example host organisms. We found that, in many cases, there is significant variation between genes

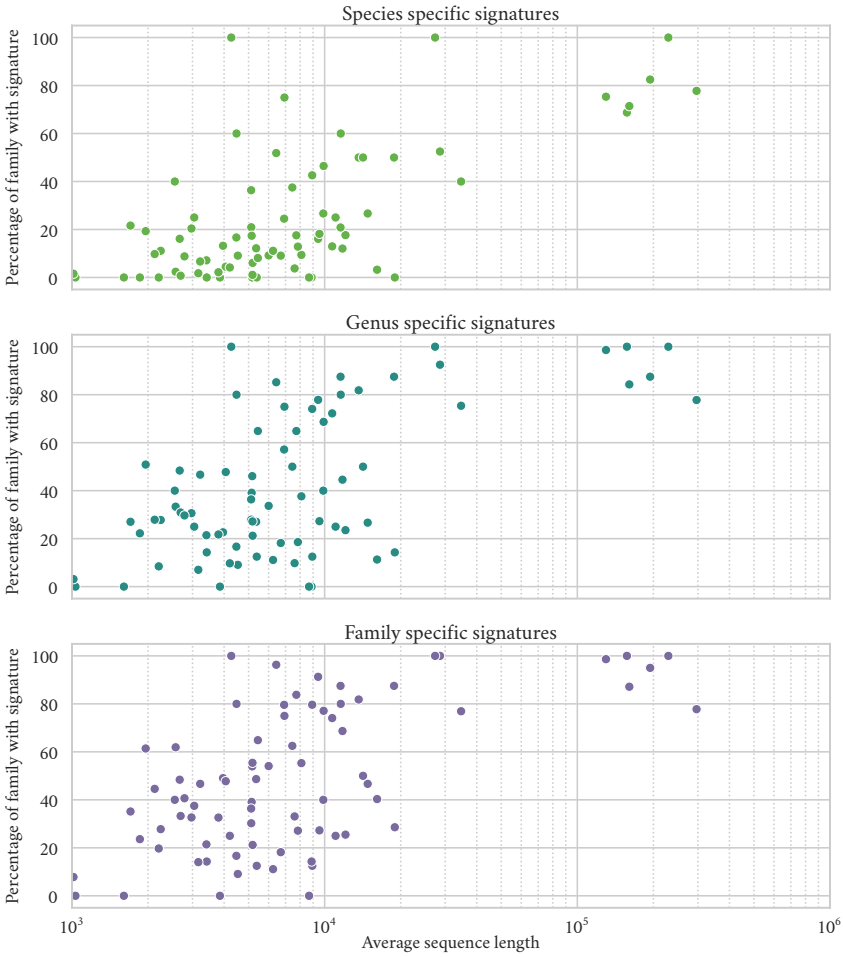

Figure 4.2 The percentage of each family that has specific signatures. Note that among the viruses with long genomes, most viruses have even species-specific signatures. Among the viruses with short signatures, the method is not able to detect specific signatures. However, for viruses with genomes around 10^4 long, there is a large variation in how many viruses from each family have specific signatures.

in a genome. As expected, we observed that this variation between genes was larger than the variation between chromosomes. In addition, the variation between genes was considerably larger than the variation between the average of the entire genomes from the same taxonomic order. By adjusting the similarity between signatures with the GC content, the signatures appear to be more conserved in the genomes of organisms than the dinucleotides, codons, and codon pairs.

Despite only a few viruses with similar signatures to their hosts, there might be other patterns of adaptation of the viruses' signatures to their host's environment. In particular, the selective pressure acting on viruses might not be similar to the selective pressure acting on the host to shape the signature. Therefore, we tested if the genomic signatures of viruses with the same host were similar. Specifically, we tested the null hypothesis that viruses with the same host were equally similar to viruses with different hosts from the same genera as the included viruses. We found that for a few hosts, viruses with that host were more similar than expected, including some of the same viruses from the virus-host analysis, namely, *Adenoviridae* and *Geminiviridae*.

In conclusion, most viruses do not have genomic signatures that are significantly adapted to be similar to their hosts' signatures. While most viruses were similar to their host's GC content, most viruses with the same hosts did not have a more similar GC content than closely related viruses with other hosts. There are a few exceptions, where a small subset of the signatures of viruses in each family appear to be adapted to be similar to other viruses' signatures with the same host or their host's signatures.

4.5 Recombination detection

To demonstrate the applicability of the specific genomic signatures, we applied them to detect recombination events. Specifically, the aim of paper V was to develop a methodology to detect the presence of recombination events in sequences. Due to the research group being interested in Herpesviruses, we used *Human alphaherpesvirus 1* (HSV-1) and *Human alphaherpesvirus 2* (HSV-2) to illustrate the approach. HSV-1 and HSV-2 interspecies recombination are relevant as HSV-1 and HSV-2 have been previously shown to be able to recombine [155], and there are circulating recombined strains [156, 157]. Such interspecies recombination can have drastic consequences on the biology of viruses [44], altough the consequences of this interspecies recombination are unknown.

We tested this approach on some artificial recombined sequences (Figure 4.3) and some previously described recombined genomes [156, 157] (Figure 4.4, Figure 4.5). These results illustrated that genomic signatures could be used to detect interspecies recombination events. The GC-corrected likelihood (Equation (3.3)) was used here, making detecting recombination events easier.

To emphasize this application, we developed a pipeline to detect recombination events in sequencing reads from patient samples. We first designed an approach to generate recombinant strains in the lab to verify the pipeline. One strain of HSV-1 and one strain of HSV-2 were grown and plaque-purified in the lab. The strains were mixed and allowed to grow at an equal multiplicity of infection (0.1) on green monkey kidney cells for 21 hours. After 21 hours, the supernatant was separated from the

cells based on the assumption that the supernatant would contain mostly complete viral particles. The DNA was extracted from the supernatant and sequenced using a nanopore minion. The green monkey kidney genome was removed while sequencing with the help of adaptive sampling.

We ran a sliding window approach with genomic signatures on the resulting reads. We found that there were reads where some parts appeared to be more similar to HSV-1 and others more similar to HSV-2. This indicates that the approach of detecting recombination events from genomic signatures is viable even when applied to reads with high error rates, such as those from long-read nanopore sequencing. However, we also note that when the recombination events involve well-characterized genomes where sequences are readily available, applying read mapping might be more efficient and accurate. Therefore, we also developed a method based on read mapping. Each read is mapped to both reference sequences and classified as recombinant if one part of the read aligns better to HSV-1 and another part aligns better to HSV-2.

As the conditions in grown lab samples and patient samples can be drastically different, with more possible contaminants in patient samples, we finally tested the approach on a random patient sample that was reported as positive for both HSV-1 and HSV-2 from the diagnostics laboratory at Sahlgrenska university hospital. Here, we additionally found reads that contained recombination events. This was done as part of methods verification and development to verify if patient samples positive for both HSV-1 and HSV-2 are positive for both strains or if they result from recombination.

In conclusion, genomic signatures can be used to detect recombination events in viruses. This is particularly useful when the parents of the recombined strains are not previously known. In addition, the approach can be used to identify prior recombination events in genomes by identifying areas where there is variation in the signature of the genome.

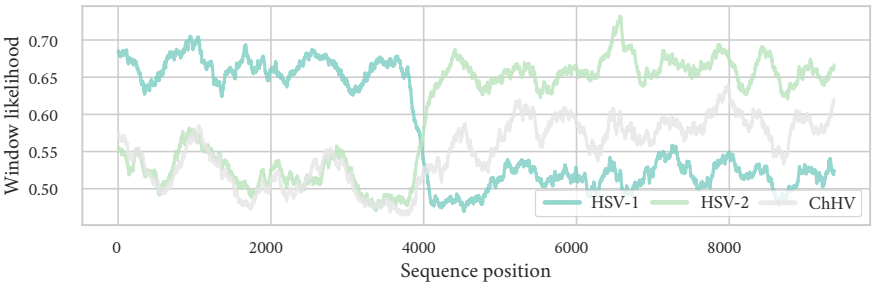

Figure 4.3 Artificial recombination between HSV-1 and HSV-2. Illustrates the sliding window scores of the recombinant. With VLMCs, it is clear that detecting the artificial recombination event is possible.

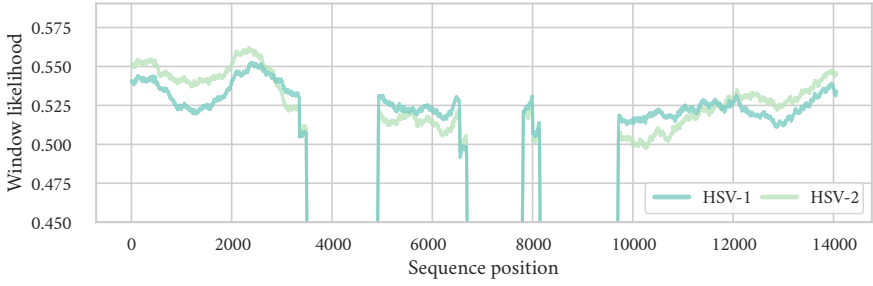

Figure 4.4 Contemporary recombination between HSV-1 and HSV-2. Illustrates the sliding window scores of the recombinant from [156]. The window likelihood drops around 4000, 7000, and 9000 nucleotides as the reference is missing these parts.

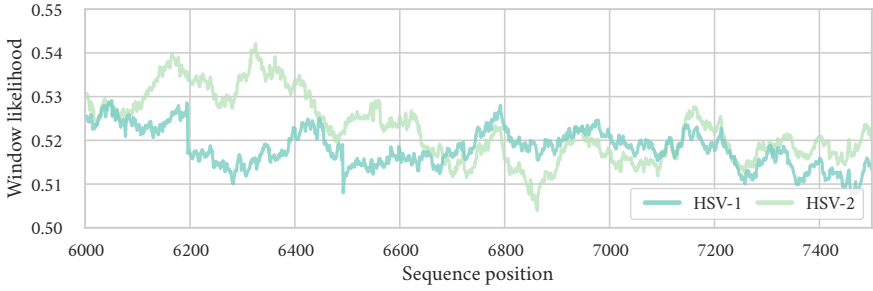

Figure 4.5 Ancient recombination between HSV-1 and HSV-2. Illustrates the sliding window scores of the recombinant from [157]. In the area between 6700 and 7100, the sequence has a higher likelihood of coming from HSV-1, while the rest of the sequence has a higher likelihood of coming from HSV-2.

Chapter 5 Discussion

By leveraging the computational advances due to our new algorithms, we analyzed genomes of viruses and hosts and found that most viruses have specific genomic signatures. These genomic signatures are likely caused by a selective pressure acting on their genomes to favor specific biological properties, such as dinucleotide and codon preferences. In addition, we have illustrated that the genomic signatures of many viruses are predominantly not adapted to be similar to the genomic signatures of their hosts other than in GC content.

5.1 Algorithmic developments

The newly developed algorithms for constructing variable-length Markov chains described in this thesis are up to 600 times faster than the previous state-of-the-art. In addition, they can operate in external memory, which allows computation on arbitrarily sized genomes. For example, it is possible to compute VLMCs on the currently largest sequenced genome [158] in less than ten minutes. While our second algorithm is faster in most cases, for smaller genomes, such as viral genomes, and where the maximal length of included k-mers is small, our first algorithm is faster (see Table 4.1).

5.1.1 New applications

These computational improvements enabled us to study the genomic signatures in viruses' hosts. However, the improvements have also expanded the types of possible analyses with VLMCs, such as large collections of genomes and raw sequencing datasets.

Example applications where VLMCs can now be applied include the study of pan-genomes. One could compute one VLMC on the core genome and separate VLMCs on the auxiliary genome. This approach might provide insights into if there are any specific patterns of core genes compared to other genes. In essence, this would be the study of pan-genome signatures. Likewise, VLMCs have previously been applied to detect horizontal gene transfer [127], which could also be applied to pan-genomes to potentially trace the origin of genes.

In addition to efficiently computing VLMCs on large genomes, our second algorithm can compute VLMCs directly on sequencing data, enabling their application to

new types of problems. In particular, we illustrated in paper II that a VLMC computed on a reference sequence is similar to a VLMC computed directly on sequencing data. These improvements enable the computation of VLMCs, for example, directly on sequencing data from patients as a means to study their general composition. Deviations from the expected VLMC in the patient (e.g., a healthy patient) would reveal a difference in which organisms are abundant and could be used to identify pathogens without classifying every individual read. Further applications can be the placement of sequenced organisms in phylogenetic trees based on whole genomes, even when there is insufficient read coverage to assemble the genomes. Even when there is sufficient data to assemble, this step could be skipped to speed up a pipeline.

5.1.2 Comparing VLMCs

We have implemented various methods of comparing VLMCs. Among them are the standard negative log-likelihood described in Section 3.1, and the d_v^* described in paper II. In addition, we have implemented one of the standard methods of comparing generative models, where sequences are generated from both models, and the negative log-likelihood is computed on each sequence with the other model [159]. There is also a d_v^* version normalized as the `CVTree` method [114] instead of as the d_2^* method [112]. The implementations of these methods are naive, with only minimal effort spent on making the comparisons computationally efficient, but there is currently some ongoing work to speed this up further.

The d_v^* is corrected for the expected similarity between two VLMCs based on the nucleotide content. Likewise, we have developed a GC-corrected negative log-likelihood measurement (Equation (3.3)). These similarity measurements ensure that the observed similarities stem from similarities in the frequent k-mers and do not purely originate from a shared GC content. Of course, when this shared GC content is of interest, this could be measured alongside the similarity, or the non-GC corrected similarities can be used. Nonetheless, we have observed that these GC-corrected similarity measurements can lead to a higher specificity.

We found that with d_v^*, the distance between genomes correlates with simulated mutation rates. Specifically, by comparing a reference *E. Coli* strain to *E. Coli* strains where we introduced some amount of mutations, the d_v^* distance increases with the number of introduced mutations (Figure 5.1). Methods using long k-mers (≥ 20) rely on estimations of the mutation rate based on the number of matching long k-mers (e.g., `mash` [109]), which have recently been improved with additional statistics on the mutation rate [160]. However, such techniques only work on long k-mers where the matching k-mers come from an evolutionary conserved sequence. We similarly observe a correlation with mutation rate, but estimating the actual mutation rate is hard because of the interdependence of k-mer counts. Instead, transforming our d_v^* to the mutation rate might be possible through regression. The advantage of our approach compared to, e.g., `mash` [109] is that the mutation rate can be estimated for higher mutation rates, where few long k-mers are shared between sequences.

As VLMCs can now be efficiently computed, they can be applied in many situations where k-mers are used today. For example, the correlation with mutation

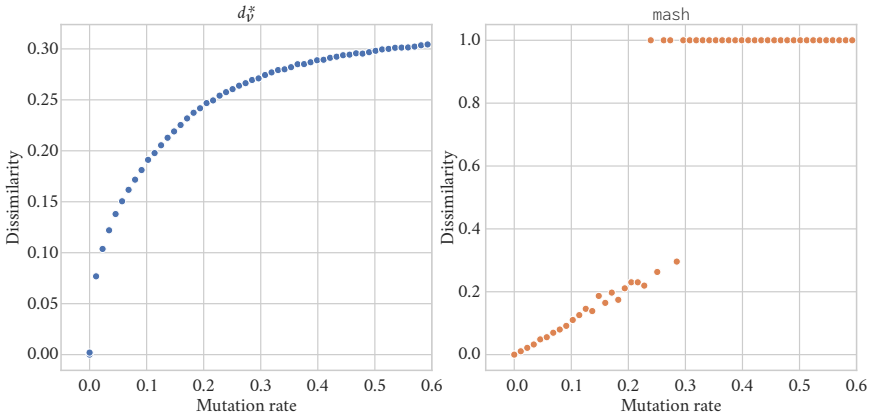

Figure 5.1 The d_v^* correlates with a simulated mutation rate. While the relationship between the mutation rate and the d_v^* dissimilarity is not linear, as with mash, it might be possible to estimate the mutation rate in contrast to mash, the d_v^* works for higher mutation rates.

rate enables them to be used for whole-genome phylogenomics, as illustrated in paper II. Likewise, VLMCs can be used to classify sequences, similar to `Kraken` [108] or `blast` [95]. This approach can be especially useful in cases where sequence homology is lacking, for example, in the study of sequences of unknown origin in metagenomic applications. In addition, considering the high mutation rates and the vast amounts of unknown viruses, this method can help provide classifications of novel viruses.

5.1.3 Model selection

The method of constructing VLMCs here has primarily focused on constraining the size of the VLMC based on the Kullback-Leibler divergence [126]. However, there are many other methods [120–125], and no available benchmark of the different methods. By analyzing the BIC [146], we have found that with the Kullback-Leibler based pruning, the optimal value of the pruning parameter Δ_{cw} (Equation (1.1)) grows with sequence size, which might indicate that this could be included in the pruning method. However, we have also observed that the resulting VLMCs are generally small with the BIC, which is generally the case for the BIC [116]. We have found these small models to work well for analyzing genomic signatures, but it appears to work less well for, e.g., the phylogenetic analysis in paper II. Likewise, both the AIC [161] and AICc [162] gave too small models for this application. In addition, the analysis of recombination events benefitted from slightly larger models than those suggested by the BIC. It is clear that the parameters of the VLMCs need to be carefully selected for each application.

5.1.4 Benefits of increased computational efficiency

Our computation advances also have environmental benefits. The carbon footprint of bioinformatics was recently reviewed [163], where the significant carbon emissions of, e.g., large-scale alignments and over-allocation of memory were highlighted. With more efficient methods, the environmental impact of evaluating methods and performing analyses is reduced, although the parallelization of methods can sometimes result in a proportionally higher amount of emissions [164].

There are also advantages for anyone running analyses based on VLMCs. As the methods run faster, it is easier to evaluate which method is appropriate for a given problem and dataset. From the perspective of ease of applying VLMCs, the most important computational achievement here was enabling analyses to run on locally available hardware, such as laptops, instead of clusters.

In addition to improving the algorithms for constructing VLMCs, considerable effort was spent on developing tools to analyze the VLMCs. Such tools include sliding-window protocols, clustering, statistical analyzes, graphical representations, and general data-analysis pipelines. These tools provide a fundamental framework for the continued analysis of genomic signatures based on VLMCs. In all cases, the methods are made available in containers to enable their portability and to provide documentation on how to build and use the methods.

5.1.5 Possible limitations

One minor limitation of our fastest algorithm of computing VLMCs, unlikely to skew analysis, is caused by the fixed size of k-mers the method uses to estimate VLMCs. Specifically, at the end of each sequence, the k-mers smaller than the maximum size are not included in the counts. For example, if the maximum length of k is 15, the 14 k-mers at the end of the sequence with length $k-1, k-2 \ldots, 1$ will not be included. For genomes, this will only have a small impact on the total counts. Note, however, that this will have a slightly larger influence on sequencing read datasets, where this small error in the counts of the smaller k-mers will influence the counts from every read. On long reads, this effect will still be small in relation to the total length size of the sequencing data. On short reads, for example, 100 nt, one k-mer will be missing per $100-k+1$ possible k-mers. However, the ends of short reads are often already of low quality, and, therefore, filtering those k-mers might not negatively influence analyses. In addition, in relation to the total number of k-mers, the missing counts make up only a small minority. However, this highlights that one should not use excessively long k-mers when training the VLMCs with this method, particularly on short reads. These same errors will be introduced for every `N` character in the sequence. Depending on the application, it might be advisable to randomly replace `N`s with random nucleotides.

One potential discrepancy between VLMCs and genomic signatures relates to selecting k-mers to include in the models. The model selection is based on which k-mers can accurately model the training sequence. As such, the included k-mers are the frequent k-mers and the k-mers where the next-symbol probabilities are informative. However, this selection of k-mers in the model does not necessarily correlate with the k-mers under selective pressure in the genome. It is possible to identify the

k-mers that currently are mutated by analyzing many contemporary strains of a virus species to identify which changes are consistently introduced (e.g., [165]). Combining this information with the genomic signatures can give a deeper understanding of how the genomes and genomic signatures evolve.

5.2 Genomic signatures in viruses

Our analysis of the genomes of viruses with VLMCs has revealed that many viruses have specific genomic signatures. In addition, these genomic signatures are often not significantly adapted to be similar to the host's signatures other than in GC content.

5.2.1 Underlying mechanisms of genomic signatures

A specific genomic signature indicates a preference for certain k-mers throughout a genome. There are many possible explanations for such a preference. Likely, they result from several factors, primarily selective pressure, but also gene duplication and replication errors.

Selective pressure

We discussed the potential mechanisms of adaptation of the nucleotide sequences of viruses that give genomic signatures in detail in paper III and introduced many possible reasons for such adaptation in Section 1.3 and Section 1.7. One likely source of the specific genomic signatures is the selective pressures that act primarily on the nucleotide sequence. These presumed selective pressures stem from the host's cellular environment and act, e.g., to optimize the translation of genes.

Among these selective pressures are those acting on the nucleotides, dinucleotides, codons, and codon pairs of a genome. For instance, a particular set of codons might be preferred to adapt the virus to the host's translational machinery. In addition, specific codon pair bias has been suggested to influence both the translation of genes and mRNA stability [86, 87]. The importance of these properties for the fitness of viruses has been demonstrated by altering the dinucleotides, codons, and codon pair usage of virus' genomes, leading to attenuated strains of the respective viruses [88, 135, 137]. In fact, even random synonymous codon changes can attenuate viruses [166]. This highlights that the specific preference of biological properties in the genomes of viruses is important for their fitness in their hosts, which results in specific genomic signatures.

Furthermore, additional and unknown selective pressures might act on viruses to shape their preference for certain k-mers, such as specific binding sites or other mechanisms involved in the replication, translation, or stability of the genome.

An alternative selective pressure acting on viral genomes might be related to the initation of recombination. A recent theory suggested that short frequent k-mers in a genome could help initiate recombination [34], leading to selective pressure on the genome to create such k-mers. This theory is based on the fact that short k-mers in stem-loop structures might help initiate the binding of different strands.

The fact that we find specific genomic signatures in many viruses reflects that there are similar selective pressures that acts on the entire genomes. Specifically, in the case where only a few genes were adapted for, e.g., high gene expression, we would not observe such a pronounced bias in the preference for a specific set of k-mers. Likely, many different selective pressures acting in tandem give rise to specific genomic signatures.

Gene duplication

Gene and genome duplication are important mechanisms of genome growth. Among viruses, such duplications are more prominent for DNA viruses than RNA viruses [41]. Immediately after a gene or whole genome duplication, there is an increase in the bias towards a particular set of k-mers, namely those present in the duplicated regions. This bias might be especially prominent when the duplicated gene is conserved, which can happen either to increase redundancy or to increase the expression of the gene product [40]. If the duplicated genes are mutated with high frequency after duplication, the duplication would not necessarily bias the genome to a specific set of preferred k-mers. However, the estimated mutation rate of viruses varies widely [25], and in some cases, even with ancient gene duplications, there might still be some remaining bias from duplications.

However, while gene duplication contributes to a bias in the preferred set of k-mers in a genome, this can not explain all of our results. Specifically, many genomes with ancient gene duplications also have distinct, specific genomic signatures. Thus, specific genomic signatures have to be caused by additional mechanisms, such as selective pressure.

Replication errors

Errors in the replication of genomes can cause neutral mutations to accumulate in genomes. In some cases, the replication machinery consistently introduces the same errors. For example, in the hepatitis C virus, the replication machinery regularly makes `A` to `G` and `C` to `U` (or vice versa) errors [167]. In humans, there is instead a tendency to remove the `CpG` dinucleotide [67]. Such errors during replication, where a specific mutation is introduced, can give rise to a pattern where certain k-mers are repeated more often than others. Given a sufficient amount of such mutations, this can give the preferred set of k-mers that we observe with the specific genomic signatures. See Figure 5.2 for an illustration.

5.2.2 Virus genomic signatures' host independence

In paper IV, we found that most viruses' genomic signatures are not similar to their hosts' genomic signatures other than in GC content. This lack of similarity of many k-mers is supported by the varied codon usage and nucleotide content of many genes in, e.g., the human genome [83]. However, we found the same pattern also for some bacteriophages, while at least some bacteria have a clear preference for a specific set of

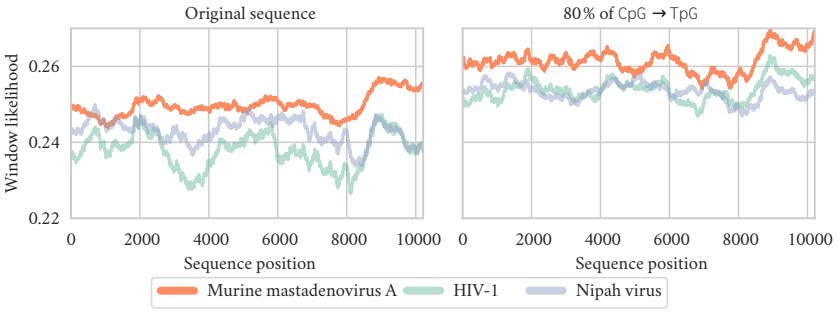

Figure 5.2 Simulation of decreased CpG dinucleotide frequency in a sequence. Illustrates the sliding window likelihood of part of *Murine adenovirus 2*, which in paper III is shown to not have a specific signature. Here, only the part of the sequence not used to train the VLMC is displayed. The genomic signature is more easily distinguishable in the presence of selective pressure to eliminate most of the CpG dinucleotides. Note that this illustration does not consider the influence on the coding sequence.

codons [168]. One possible explanation for this discrepancy is that the genomic signatures in viruses are caused by different selection pressure from the selection pressure acting on the host's genome. For example, some viruses carry t-RNA-like genes [5] or are, for other reasons, less dependent on the specifics of the host's gene expression machinery [169]. Therefore, the most efficient codons in the host might not be the most efficient for the virus.

There were a few exceptions where the viruses had similar signatures to their hosts and similar signatures to other viruses with the same host. These included viruses from the dsDNA family *Adenoviridae* and the ssRNA(-) family *Coronaviridae*. For *Adenoviridae*, all viruses that infect *Homo sapiens*, *Macaca fascicularis*, *Macaca mulatta*, *Chlorocebus aethiops*, *Gallus gallus*, and *Bos taurus* were respectively more similar than expected. For *Coronaviridae*, all viruses that infect the bat families *Pteropodidae*, *Rhinolophidae*, and *Vespertilionidae* were respectively more similar than expected. However, we note that only a small fraction of the viruses from the families were similar to a host, which might indicate that such adaptation is not prominent. Instead, this might indicate that all viruses with those hosts are more closely phylogenetically related than the other viruses in the same genera. This would be expected when the viruses speciated through intra-host speciation, where multiple virus species evolved in the same host.

Similarly to previous results [133], we found that many viruses had a similar GC content to their hosts. However, only a few viruses with the same host were significantly similar in GC content. One explanation is that the GC content of many related viruses is caused by a shared selective pressure acting on these genomes that is not significantly different between different hosts, such as when infecting similar niches in different hosts. Furthermore, we found that the bacteriophages were the group of viruses most similar to their hosts in GC content, which might indicate that this is an important adaptation for phages. This pronounced similarity in GC con-

tent in conjunction with short k-mers has been previously used to predict the host of some phages [113, 142]. Note, however, that the lack of significant similarity in the GC-corrected similarity might indicate that such predictions need to be considered carefully.

5.2.3 Evolution in multiple hosts

Additionally, the patterns of genomic signatures in viruses are likely the result of selection pressures acting on viruses over their and their ancestors' evolution. This is specifically important when the ancestral host of the virus is dissimilar from its current host. Both speciation through host switches and co-speciation places viruses under different selective pressure than before speciation. Therefore, the genomic signatures may reflect selective pressure both from an ancient host as well as the current host of the virus.

Furthermore, some viruses infect several hosts. Many viruses use vectors to move from one host to another, for instance, *Flaviviridae* and *Geminiviridae*. It has been suggested for some viruses that the virus is primarily adapted to the non-vector host [170] due to an observed dinucleotide bias. In addition, some viruses are generalists and infect multiple species [171]. In both cases, selective pressure might be acting on the viral genomes from multiple hosts, and therefore, there is no significant adaptation to one specific host's genomic signature.

5.2.4 Variations in specific genome-wide signatures

The method used in paper III could not detect genomic signatures in some viruses. Here, we illustrated that this was partly an issue with sequence length, as subsequences of 5000 nt from viruses with species-specific signatures often did not present specific signatures. Thus, there is a methodological limitation in the detection of genomic signatures for shorter genomes.

However, more of those subsequences had a specific signature than the viruses with genomes between 5000 nt to $10\,000$ nt long. This might indicate that, for some viruses, there might not be pronounced genomic signatures. To illustrate this, consider the normalized likelihood of sliding windows in four viruses (Figure 5.3). As in paper III, the signatures are computed only on the last $70\,\%$ of the sequence. For viruses with specific signatures, we expect to observe a high likelihood for all windows, including those in the first $30\,\%$ of the sequence, which is not used to train the signatures. If all windows have a high likelihood, this indicates a specific signature and thus a genome-wide preference for certain k-mers.

For viruses with no specific genomic signatures, there are two possibilities. (i) There are no particular preferences for certain k-mers, so the signature is difficult to distinguish from other viruses. This lack of preference can be seen for the *Tobacco mosaic virus* in Figure 5.3, with an average window likelihood below 0.26, and many areas are difficult to distinguish from the *Tomato mottle virus*. Compare this to HSV-1 in Figure 5.3, which has an average window likelihood above 0.26 and a large difference in likelihood to other viruses. (ii) There are preferences for specific k-mers, but they vary considerably within the sequence. This case is exemplified by the *Murid*

betaherpesvirus 2 in Figure 5.3, where the middle of the sequence has an average likelihood close to 0.28, but the beginning and end have considerably lower likelihoods. Such variation in the genomic signature can also be seen for the *Murine mastadenovirus A* in Figure 5.3, where the first 30 % of the sequence has considerably lower average likelihood than the 70 % of sequence used to train the model.

There are likely both viruses with no detectable signatures due to methodological limitations, but there are also viruses that lack specific signatures. This lack of specific signatures might stem from a mix of several different signatures in the genomes or simply no pronounced preference for any specific k-mers.

5.3 Recombination and genomic signatures

We illustrated an application of genomic signatures on the identification of recombinants. Specifically, we implemented a method of detecting interspecies recombination of HSV-1 and HSV-2. These recombinants were generated *in silico* and *in vitro*.

The *in vitro* recombinants were harvested from one strain each of HSV-1 and HSV-2 that were simultaneously inoculated on green monkey kidney cells. While we harvested the supernatant of these cells with the assumption that there would be more complete virions in the supernatant, the supernatant could also contain genome fragments from lysed cells. Therefore, further tests are needed to ensure that the predicted recombined sequences come from viable viruses.

Furthermore, the lab-grown strains were harvested after only 21 hours to prevent either strain from out-competing the other. Specifically, the HSV-2 strain, in this case, grows faster than the HSV-1 strain. While we found recombinant reads after 21 hours, it might be possible to increase our understanding of how quickly and readily the viruses recombine by analyzing the supernatant at multiple time periods.

When the suspected recombination event involves two organisms for which there are available sequences, applying homology-based tools, such as read-mapping, might give better results than the genomic signatures. As pangenomes and pan-genome graphs become available for herpesviruses, an even better option might be to use a pan-genomics mapper to directly identify which parts of the read are the most similar to either strain.

However, in cases where there are no known homologous sequences or the specific species involved in the recombination is not known, a sliding window analysis of the sequencing reads based on genomic signatures can identify areas where the signature differs from the expected signature. These areas might correspond to horizontal gene transfer, such as through recombination.

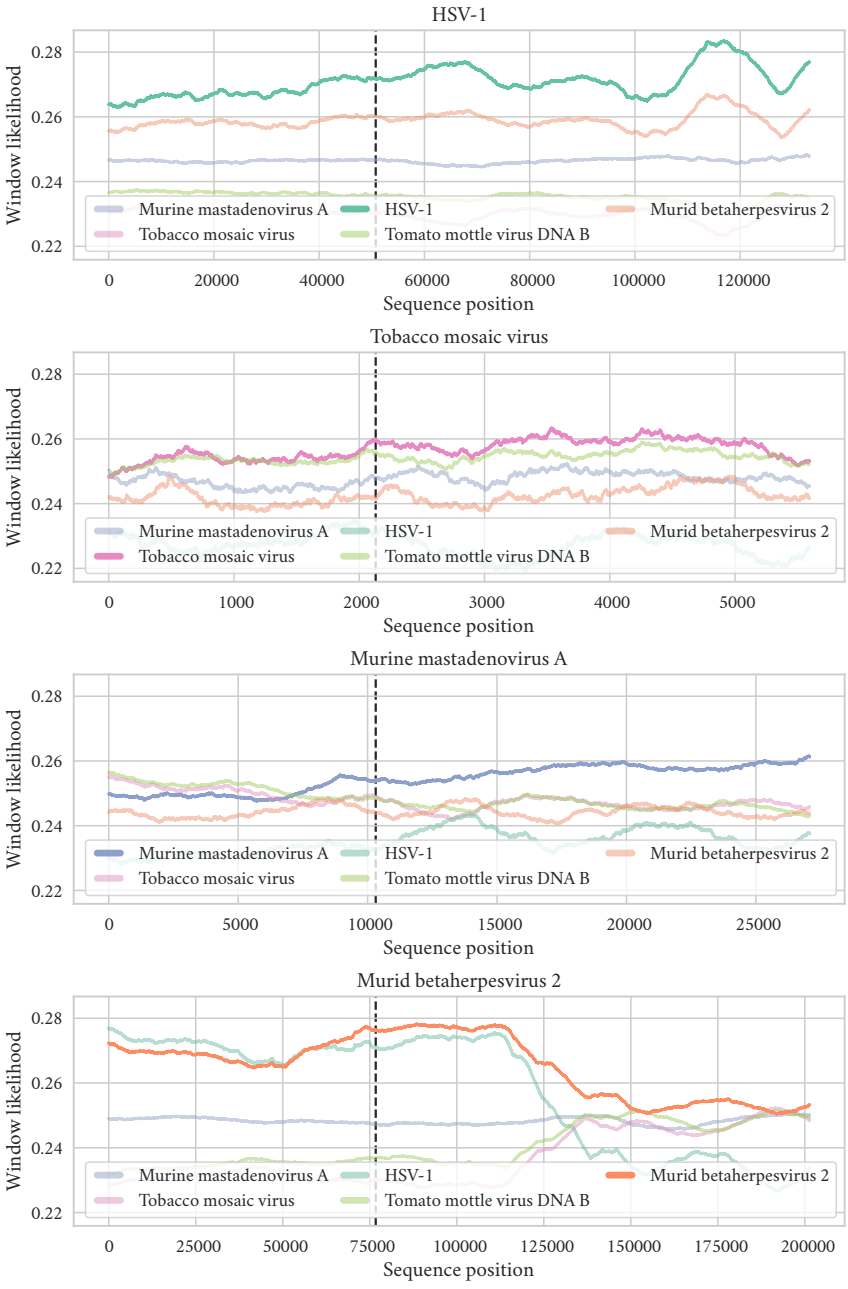

Figure 5.3 Normalized sliding window likelihood of four example viruses. Shows HSV-1 which has a species-specific signature and *Tobacco mosaic virus*, *Murine mastadenovirus A*, and *Murid betaherpesvirus 2* where the method could not detect a specific signatures. The VLMCs are computed on the sequence to the right of the dotted vertical line.

Chapter 6 Outlook

We have shown that many viruses have specific genomic signatures, likely caused by evolutionary selection on the genome. What these selection pressures are, however, has yet to be discovered. With the knowledge that we can detect specific genomic signatures in viruses, future work could analyze each selection pressure individually. Specifically, it might be possible to determine in detail which selective pressures influence the genomic signatures of certain viral species, and to which degree. It is already clear that the GC content significantly impacts the signatures. Thus, selection pressures acting on the GC content influence the signature. By specifically analyzing, for example, dinucleotide, codon, and codon pair usage, in relation to which viruses have specific genomic signatures, it might be possible to determine their specific influence on the genomic signatures. Likewise, we could alter, e.g., the codon bias of a virus genome to study the impact on the specific genomic signature.

The presence of specific genomic signatures in viruses indicates a link with the fitness of the virus. A potential application of this is to introduce changes in a genome that contradict the genomic signature. This approach is, in principle, similar to codon and codon pair deoptimization techniques [88, 134]. As the genomic signatures include information from multiple sources, such as dinucleotides, codons, and codon pairs, it might be possible to achieve a higher level of attenuation than with any individual approach. In addition, the flexible nature of the signatures also makes this approach adaptive to the specific virus. As such, the genomic signatures could be used to attenuate viruses to generate potential vaccine candidates.

Likewise, it might be possible to adapt a gene to a new host by changing the gene to be similar to the genomic signature of the rest of the genome, for example, for use with viral vectors. However, as we have noticed that many viruses are not similar in signature to their hosts other than in GC content, optimizing the genomic signatures of viral genes towards a host signature might not increase protein production. In particular, based on the assumption that viruses are already highly adapted to their environments, it is unlikely that such an approach could create a more fit virus in its natural environment.

By studying the genomic signatures of individual viruses and families, it might be possible to gain insights into their biology. For instance, the sliding window analysis of genomes can reveal regions of genomes where the selective pressure differs from other regions. Such regions might indicate horizontal gene transfer, e.g., as illustrated here, or other areas of biological importance, including hypervariable regions.

The presence of genomic signatures indicates the potential for approaches that can classify unknown sequences. With the increase in computational efficiency presented here, the genomic signatures can be applied to modern datasets, e.g., metagenomic datasets and pangenomes. The genomic signatures can identify more remotely related species than the long unique k-mers that, e.g., `Kraken 2` [108] uses, or the homologous regions that `BLAST` [95] uses. As such, genomic signatures can be used to classify reads that those methods can not. This represents a homology-free method that could help shed light on the vast amounts of genomic dark matter by providing putative classification. Our method can be especially important for immunocompromised patients with opportunistic infections, where a putative assignment of an unknown pathogen can help suggest treatments.

The importance of systems that recognize and classify viruses has recently become apparent. Such systems will only increase in importance as many new viruses emerge, both by being released as ancient ice melts [172] and from pathogens that recently have started to migrate as their vectors can survive in areas where they previously could not [173]. Our methods could be used to monitor samples for viruses, even when such viruses lack sequence homology to known viruses.

As the estimated number of viruses is in the millions, but the officially recognized viruses are only $10\,434$, there is a need for novel methods to help characterize viruses. The genomic signatures and methods presented here offer a promising approach to detecting viruses based on biological properties and not sequence homology and can help identify emerging viruses.

Acknowledgements

I want to thank everyone who has helped, guided and supported me during these years of Ph.D. studies. Foremost I would like to thank my supervisor Peter Norberg, whose scientific guidance, help in the field of virology, and bridging of biology and computer science have been instrumental during these years. I would equally like to thank my co-supervisor Alexander Schliep, who has spent considerably more time than could be expected of any co-supervisor to guide the computational advances and general analyses and to help me better grasp the field of bioinformatics. My fellow Ph.D. student in the group, Martin Holmudden, for interesting talks about virology and bioinformatics and for valuable help in developing the ideas in this thesis. Thanks to Adina Lupu for helping me understand biology during the early days of my Ph.D. Thanks also to Yann Bertrand and Scott V. Edwards for assisting with their research expertise.

Of course, I would also like to thank everyone in the virology department who has helped with various tasks, been encouraging, and generally created a pleasant atmosphere. In particular, thanks to Maria Johansson for initially showing me around the virology building and for instructing me in wet lab work. Thanks to Josefin Olausson, Maria Andersson, Johan Ringlander, Joakim Bedner Stenbäck, and Daniel Schmidt for invaluable help with nanopore sequencing and DNA extraction.

Also, all the Msc students I've supervised, Sebastian Norlin, Jan Qvick-Wester, Frans Wallin, Johan Atterfors, Sebastian Holm, Erik Söderpalm, and Filip Helmroth, thank you for spending time helping us advance our research.

Thanks to the Ph.D. students, teachers, and organizers of the MedBioInfo research school for valuable insights into many aspects of bioinformatics.

Thanks also to Erik Norlander, with whom I completed the MSc project that led to this thesis. Thanks to all my other friends from Chalmers, especially André Samuelsson, Mattias Nilsen, Johan Lindskogen, and Ivar Josefsson, for their friendship and support in all things software engineering. Thanks to my once-roommate Martin Tomasson for the enthusiasm and experiences. Thanks to Niklas Vågstedt for being the source of my love of programming and Phoebe Jönsson for the enthusiasm and friendship.

I would also like to thank my family, Elias, Erik, Kristina, Bengt, and Barbro, for all their unwavering support and encouragement.

Finally, I would like to thank Josefina Andreasson for your continued patience and support, for always helping with everything, and for reminding me to take a break. Without you, this thesis would not have been possible.

Bibliography

[1] Alison A. McBride. "Mechanisms and strategies of papillomavirus replication". In: *Biological Chemistry* 398 (8 July 2017), pp. 919–927. DOI: 10.1515/hsz-2017-0113.

[2] Naomi Kitamura et al. "Primary structure, gene organization and polypeptide expression of poliovirus RNA". In: *Nature* 291.5816 (1981), pp. 547–553. DOI: 10.1038/291547a0.

[3] S. K. Weller and D. M. Coen. "Herpes Simplex Viruses: Mechanisms of DNA Replication". In: *Cold Spring Harbor Perspectives in Biology* 4 (9 Sept. 2012), a013011–a013011. DOI: 10.1101/cshperspect.a013011.

[4] B. Moss. "Poxvirus DNA Replication". In: *Cold Spring Harbor Perspectives in Biology* 5 (9 Sept. 2013), a010199–a010199. DOI: 10.1101/cshperspect.a010199.

[5] Theo W Dreher. "Viral tRNAs and tRNA-like structures". In: *Wiley Interdisciplinary Reviews: RNA* 1.3 (2010), pp. 402–414. DOI: 10.1002/wrna.42.

[6] Thomas C. Mettenleiter. "Budding events in herpesvirus morphogenesis". In: *Virus Research* 106 (2 Dec. 2004), pp. 167–180. DOI: 10.1016/j.virusres.2004.08.013.

[7] G Chinnadurai. "Control of apoptosis by human adenovirus genes". In: *Seminars in VIROLOGY*. Vol. 8. 5. Elsevier. 1998, pp. 399–408. DOI: 10.1006/smvy.1997.0139.

[8] Christian Münz. "The Autophagic Machinery in Viral Exocytosis". In: *Frontiers in Microbiology* 8 (FEB Feb. 2017). DOI: 10.3389/fmicb.2017.00269.

[9] Evelyne Richet, Peter Abcarian, and Howard A Nash. "Synapsis of attachment sites during lambda integrative recombination involves capture of a naked DNA by a protein-DNA complex". In: *Cell* 52.1 (1988), pp. 9–17. DOI: 10.1016/0092-8674(88)90526-0.

[10] Christophe Marchand et al. "Mechanisms and inhibition of HIV integration". In: *Drug Discovery Today: Disease Mechanisms* 3 (2 June 2006), pp. 253–260. DOI: 10.1016/j.ddmec.2006.05.004.

[11] Bjørn Grinde. "Herpesviruses: latency and reactivation–viral strategies and host response". In: *Journal of oral microbiology* 5.1 (2013), p. 22766. DOI: 10.3402/jom.v5i0.22766.

[12] Dmitri Ivanowski. "Ueber die mosaikkrankheit der tabakspflanze". In: *St Petersb Acad Imp Sci Bul* 35 (1892), pp. 67–70.

[13] MW Beijerinck. "Concerning a contagium viwm fluidum as cause of the spot disease of tobacco leaves". In: *Phytopathol Class* 7.1 (1898), pp. 33–52.

[14] Jens H. Kuhn et al. "Classify viruses - the gain is worth the pain". In: *Nature* (2019), pp. 8–10. DOI: 10.1007/s00705-018-04136-2.

[15] Peter J Walker et al. "Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022)". In: *Archives of virology* 167.11 (2022), pp. 2429–2440. DOI: 10.1007/s00705-022-05516-5.

[16] Jennifer Louten. "Virus structure and classification". In: *Essential human virology*. Elsevier, 2016, pp. 19–29. DOI: 10.1016/B978-0-12-800947-5.00002-8.

[17] D Baltimore. "Expression of animal virus genomes". In: *Bacteriological Reviews* 35 (3 Sept. 1971), pp. 235–241. DOI: 10.1128/br.35.3.235-241.1971.

[18] Eugene V Koonin, Mart Krupovic, and Vadim I Agol. "The Baltimore classification of viruses 50 years later: how does it stand in the light of virus evolution?" In: *Microbiology and Molecular Biology Reviews* 85.3 (2021), e00053–21. DOI: 10.1128/MMBR.00053-21.

[19] Brian M Meehan et al. "Sequence of porcine circovirus DNA: affinities with plant circoviruses". In: *Journal of General Virology* 78.1 (1997), pp. 221–227. DOI: 10.1099/0022-1317-78-1-221.

[20] Nadège Philippe et al. "Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes". In: *Science* 341 (6143 July 2013), pp. 281–286. DOI: 10.1126/science.1239181.

[21] Gita Mahmoudabadi and Rob Phillips. "A comprehensive and quantitative exploration of thousands of viral genomes". In: *Elife* 7 (2018), e31955.

[22] Igor B Rogozin et al. "Congruent evolution of different classes of non-coding DNA in prokaryotic genomes". In: *Nucleic acids research* 30.19 (2002), pp. 4264–4271. DOI: 10.1093/nar/gkf549.

[23] John S Mattick. "RNA regulation: a new genetics?" In: *Nature Reviews Genetics* 5.4 (2004), pp. 316–323.

[24] International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome". In: *Nature* 431 (7011 Oct. 2004), pp. 931–945. DOI: 10.1038/nature03001.

[25] Rafael Sanjuán and Pilar Domingo-Calap. "Mechanisms of viral mutation". In: *Cellular and Molecular Life Sciences* 73 (23 Dec. 2016), pp. 4433–4448. DOI: 10.1007/s00018-016-2299-6.

[26] Charles Darwin. *On the origin of species*. John Murray, 1859.

[27] Gregor Mendel. *Versuche über Pflanzenhybriden*. Verhandlungen des naturforschenden Vereines in Brünn, 1866.

[28] Motoo Kimura. "Evolutionary rate at the molecular level". In: *Nature* 217 (1968), pp. 624–626.

[29] David Sadava et al. *Life: the science of biology*. Vol. 11. Macmillan, 2016, p. 432.

[30] John T McCrone and Adam S Lauring. "Genetic bottlenecks in intraspecies virus transmission". In: *Current opinion in virology* 28 (2018), pp. 20–25. DOI: 10.1016/j.coviro.2017.10.008.

[31] Duncan J McGeoch et al. "Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses". In: *Journal of molecular biology* 247.3 (1995), pp. 443–458. DOI: 10.1006/jmbi.1995.0152.

[32] Anderson F Brito et al. "Intrahost speciations and host switches played an important role in the evolution of herpesviruses". In: *Virus Evolution* 7 (1 Jan. 2021). DOI: 10.1093/ve/veab025.

[33] Enosh Tomer et al. "Coalescing replication compartments provide the opportunity for recombination between coinfecting herpesviruses". In: *FASEB Journal* 33 (8 Aug. 2019), pp. 9388–9403. DOI: 10.1096/fj.201900032R.

[34] Donald R Forsdyke. "Success of alignment-free oligonucleotide (k-mer) analysis confirms relative importance of genomes not genes in speciation and phylogeny". In: *Biological Journal of the Linnean Society* 128 (July 2019), pp. 239–250. DOI: 10.1093/biolinnean/blz096.

[35] Paul M. Sharp and Beatrice H. Hahn. "The evolution of HIV-1 and the origin of AIDS". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1552 Aug. 2010), pp. 2487–2494. DOI: 10.1098/rstb.2010.0031.

[36] José M Malpica et al. "The Rate and Character of Spontaneous Mutation in an RNA Virus". In: *Genetics* 162 (4 Dec. 2002), pp. 1505–1511. DOI: 10.1093/genetics/162.4.1505.

[37] José M. Cuevas et al. "Human norovirus hyper-mutation revealed by ultra-deep sequencing". In: *Infection, Genetics and Evolution* 41 (July 2016), pp. 233–239. DOI: 10.1016/J.MEEGID.2016.04.017.

[38] Pakorn Aiewsakun and Aris Katzourakis. "Time-Dependent Rate Phenomenon in Viruses". In: *Journal of Virology* 90 (16 Aug. 2016). Ed. by S. R. Ross, pp. 7184–7195. DOI: 10.1128/JVI.00593-16.

[39] Peter Simmonds, Pakorn Aiewsakun, and Aris Katzourakis. "Prisoners of war — host adaptation and its constraints on virus evolution". In: *Nature Reviews Microbiology* 17 (5 May 2019), pp. 321–328. DOI: 10.1038/s41579-018-0120-2.

[40] Jianzhi Zhang. "Evolution by gene duplication: an update". In: *Trends in ecology & evolution* 18.6 (2003), pp. 292–298. DOI: 10.1016/S0169-5347(03)00033-8.

[41] Etienne Simon-Loriere and Edward C. Holmes. "Gene Duplication Is Infrequent in the Recent Evolutionary History of RNA Viruses". In: *Molecular Biology and Evolution* 30 (6 June 2013), pp. 1263–1269. DOI: 10.1093/molbev/mst044.

[42] Yuxia Gao et al. "Extent and evolution of gene duplication in DNA viruses". In: *Virus Research* 240 (Aug. 2017), pp. 161–165. DOI: 10.1016/j.virusres.2017.08.005.

[43] Karsten Suhre. "Gene and Genome Duplication in Acanthamoeba polyphaga Mimivirus". In: *Journal of Virology* 79 (22 Nov. 2005), pp. 14095–14101. DOI: 10.1128/jvi.79.22.14095-14101.2005.

[44] Marcos Pérez-Losada et al. "Recombination in viruses: mechanisms, methods of study, and evolutionary consequences". In: *Infection, Genetics and Evolution* 30 (2015), pp. 296–307.

[45] Xiaojun Li et al. "Emergence of SARS-CoV-2 through recombination and strong purifying selection". In: *Science Advances* 6 (27 July 2020), eabb9153. DOI: 10.1126/sciadv.abb9153.

[46] Juan Ángel Patiño-Galindo et al. "Recombination and lineage-specific mutations linked to the emergence of SARS-CoV-2". In: *Genome Medicine* 13.1 (2021), pp. 1–14. DOI: 10.1186/s13073-021-00943-6.

[47] Pierre Lefeuvre and Enrique Moriones. "Recombination as a motor of host switches and virus emergence: geminiviruses as case studies". In: *Current Opinion in Virology* 10 (2015), pp. 14–19. DOI: 10.1016/j.coviro.2014.12.005.

[48] Alltalents T Murahwa, Mqondisi Tshabalala, and Anna-Lise Williamson. "Recombination between high-risk human papillomaviruses and non-human primate papillomaviruses: evidence of ancient host switching among alphapapillomaviruses". In: *Journal of Molecular Evolution* 88 (2020), pp. 453–462. DOI: 10.1007/s00239-020-09946-0.

[49] H.J. Muller. "The relation of recombination to mutational advance". In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 1 (1 May 1964), pp. 2–9. DOI: 10.1016/0027-5107(64)90047-8.

[50] R G Webster et al. "Evolution and ecology of influenza A viruses". In: *Microbiological Reviews* 56 (1 Mar. 1992), pp. 152–179. DOI: 10.1128/mr.56.1.152-179.1992.

[51] David K Clarke et al. "The red queen reigns in the kingdom of RNA viruses". In: *Proceedings of the National Academy of Sciences* 91.11 (1994), pp. 4821–4824. DOI: 10.1073/pnas.91.11.4821.

[52] Marilyn J. Roossinck. "Move Over, Bacteria! Viruses Make Their Mark as Mutualistic Microbial Symbionts". In: *Journal of Virology* 89 (13 July 2015). Ed. by S. Schultz-Cherry, pp. 6532–6535. DOI: 10.1128/JVI.02974-14.

[53] Marilyn J. Roossinck. "The good viruses: viral mutualistic symbioses". In: *Nature Reviews Microbiology* 9 (2 Feb. 2011), pp. 99–108. DOI: 10.1038/nrmicro2491.

[54] "The healthy human virome: from virus–host symbiosis to disease". In: *Current Opinion in Virology* 47 (Apr. 2021), pp. 86–94. DOI: 10.1016/j.coviro.2021.02.002.

[55] Pierre Lefeuvre et al. "Evolution and ecology of plant viruses". In: *Nature Reviews Microbiology* 17.10 (2019), pp. 632–644. DOI: 10.1038/s41579-019-0232-3.

[56] Peter J Kerr et al. "Evolutionary history and attenuation of myxoma virus on two continents". en. In: *PLoS Pathogens* 8.10 (2012), e1002950. DOI: 10.1371/journal.ppat.1002950.

[57] Gary L Disbrow et al. "Codon optimization of the HPV-16 E5 gene enhances protein expression". In: *Virology* 311.1 (2003), pp. 105–114. DOI: 10.1016/S0042-6822(03)00129-6.

[58] F. Sanger et al. "Nucleotide sequence of bacteriophage ΦX174 DNA". In: *Nature* 265 (5596 Feb. 1977), pp. 687–695. DOI: 10.1038/265687a0.

[59] B. G. Barrell, G. M. Air, and C. A. Hutchison. "Overlapping genes in bacteriophage ΦX174". In: *Nature* 264 (5581 Nov. 1976), pp. 34–41. DOI: 10.1038/264034a0.

[60] Peter Yakovchuk, Ekaterina Protozanova, and Maxim D. Frank-Kamenetskii. "Base-stacking and base-pairing contributions into thermal stability of the DNA double helix". In: *Nucleic Acids Research* 34 (2 Feb. 2006), pp. 564–574. DOI: 10.1093/nar/gkj454.

[61] Laurence D Hurst and Alexa R Merchant. "High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268.1466 (2001), pp. 493–497. DOI: 10.1098/rspb.2000.1397.

[62] Konrad U. Foerstner et al. "Environments shape the nucleotide composition of genomes". In: *EMBO Reports* 6 (12 Dec. 2005), pp. 1208–1213. DOI: 10.1038/sj.embor.7400538.

[63] International Human Genome Sequencing Consortium. "Initial sequencing and analysis of the human genome". In: *Nature* 409 (6822 Feb. 2001), pp. 860–921. DOI: 10.1038/35057062.

[64] Laurent Duret, Dominique Mouchiroud, and Christian Gautier. "Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores". In: *Journal of molecular evolution* 40 (1995), pp. 308–317. DOI: 10.1007/BF00163235.

[65] Chris Burge and Samuel Karlin. "Prediction of complete gene structures in human genomic DNA". In: *Journal of Molecular Biology* 268 (1 Apr. 1997), pp. 78–94. DOI: 10.1006/jmbi.1997.0951.

[66] Winston Salser. "Globin mRNA sequences: analysis of base pairing and evolutionary implications". In: *Cold Spring Harbor Symposia on Quantitative Biology*. Vol. 42. Cold Spring Harbor Laboratory Press. 1978, pp. 985–1002. DOI: 10.1101/SQB.1978.042.01.099.

[67] Adrian P Bird. "DNA methylation and the frequency of CpG in animal DNA". In: *Nucleic acids research* 8.7 (1980), pp. 1499–1504. DOI: 10.1093/nar/8.7.1499.

[68] S Karlin, W Doerfler, and L R Cardon. "Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses?" In: *Journal of Virology* 68 (5 1994), pp. 2889–2897. DOI: 10.1128/jvi.68.5.2889-2897.1994.

[69] Benjamin D. Greenbaum et al. "Patterns of evolution and host gene mimicry in influenza and other RNA viruses". In: *PLoS Pathogens* 4 (6 2008), pp. 1–9. DOI: 10.1371/journal.ppat.1000079.

[70] Denise Lecossier et al. "Hypermutation of HIV-1 DNA in the absence of the Vif protein". In: *Science* 300.5622 (2003), pp. 1112–1112. DOI: 10.1126/science.1083338.

[71] Bastien Mangeat et al. "Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts". In: *Nature* 424.6944 (2003), pp. 99–103. DOI: 10.1038/nature01709.

[72] Hui Zhang et al. "The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA". In: *Nature* 424.6944 (2003), pp. 94–98. DOI: 10.1038/nature01707.

[73] Matthew A. Takata et al. "CG dinucleotide suppression enables antiviral defence targeting non-self RNA". In: *Nature* 550 (7674 2017), pp. 124–127. DOI: 10.1038/nature24039.

[74] S Karlin and C Burge. "Dinucleotide relative abundance extremes: a genomic signature". In: *Trends in Genetics* 11 (7 July 1995), pp. 283–290. DOI: 10.1016/S0168-9525(00)89076-9.

[75] Allan Campbell, Jan Mrázek, and Samuel Karlin. "Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA". In: *Proceedings of the National Academy of Sciences* 96 (16 Aug. 1999), pp. 9184–9189. DOI: 10.1073/pnas.96.16.9184.

[76] Paul M. Sharp and Wen Hsiung Li. "The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications". In: *Nucleic Acids Research* 15 (3 1987), pp. 1281–1295. DOI: 10.1093/nar/15.3.1281.

[77] SG Andersson and CG Kurland. "Codon preferences in free-living microorganisms". In: *Microbiological reviews* 54.2 (1990), pp. 198–210. DOI: 10.1128/mr.54.2.198-210.1990.

[78] Joshua B. Plotkin and Grzegorz Kudla. "Synonymous but not the same: The causes and consequences of codon bias". In: *Nature Reviews Genetics* 12 (1 2011), pp. 32–42. DOI: 10.1038/nrg2899.

[79] Toshimichi Ikemura. "Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system". In: *Journal of molecular biology* 151.3 (1981), pp. 389–409. DOI: 10.1016/0022-2836(81)90003-6.

[80] Toshimichi Ikemura. "Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs". In: *Journal of molecular biology* 158.4 (1982), pp. 573–597. DOI: 10.1016/0022-2836(82)90250-9.

[81] Swaine L. Chen et al. "Codon usage between genomes is constrained by genome-wide mutational processes". In: *Proceedings of the National Academy of Sciences* 101 (10 Mar. 2004), pp. 3480–3485. DOI: 10.1073/pnas.0307827100.

[82] Laura A. Shackelton, Colin R. Parrish, and Edward C. Holmes. "Evolutionary Basis of Codon Usage and Nucleotide Composition Bias in Vertebrate DNA Viruses". In: *Journal of Molecular Evolution* 62 (5 May 2006), pp. 551–563. DOI: 10.1007/s00239-005-0221-1.

[83] Paul M Sharp et al. "Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity". In: *Nucleic Acids Research* 16 (17 1988), pp. 8207–8211. DOI: 10.1093/nar/16.17.8207.

[84] Joshua B Plotkin, Harlan Robins, and Arnold J Levine. "Tissue-specific codon usage and the expression of human genes". In: *Proceedings of the National Academy of Sciences* 101.34 (2004), pp. 12588–12591. DOI: 10.1073/pnas.0404957101.

[85] George A Gutman and G Wesley Hatfield. *Nonrandom utilization of codon pairs in Escherichia coli*. 1989, pp. 3699–3703. DOI: 10.1073/pnas.86.10.3699.

[86] Caitlin E Gamble et al. "Adjacent codons act in concert to modulate translation efficiency in yeast". In: *Cell* 166.3 (2016), pp. 679–690. DOI: 10.1016/j.cell.2016.05.070.

[87] Nicole Groenke et al. "Mechanism of virus attenuation by codon pair deoptimization". In: *Cell reports* 31.4 (2020), p. 107586. DOI: 10.1016/j.celrep.2020.107586.

[88] J. Robert Coleman et al. "Virus attenuation by genome-scale changes in codon pair bias". In: *Science* 320 (5884 2008), pp. 1784–1787. DOI: 10.1126/science.1155761.

[89] Margaret Oakley Dayhoff. "Computer analysis of protein evolution". In: *Scientific American* 221.1 (1969), pp. 86–95.

[90] M Dayhoff, R Schwartz, and B Orcutt. "A model of evolutionary change in proteins". In: *Atlas of protein sequence and structure* 5 (1978), pp. 345–352.

[91] Sean R Eddy. "Where did the BLOSUM62 alignment score matrix come from?" In: *Nature biotechnology* 22.8 (2004), pp. 1035–1036. DOI: 10.1038/nbt0804-1035.

[92] Saul B Needleman and Christian D Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In: *Journal of molecular biology* 48.3 (1970), pp. 443–453. DOI: 10.1016/0022-2836(70)90057-4.

[93] Christine N Tennyson, Henry J Klamut, and Ronald G Worton. "The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced". In: *Nature genetics* 9.2 (1995), pp. 184–190. DOI: 10.1038/ng0295-184.

[94] Maria Chatzou et al. "Multiple sequence alignment modeling: methods and applications". In: *Briefings in bioinformatics* 17.6 (2016), pp. 1009–1023. DOI: /10.1093/bib/bbv099.

[95] Stephen F Altschul et al. "Basic local alignment search tool". In: *Journal of molecular biology* 215.3 (1990), pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2.

[96] Heng Li. "Minimap2: pairwise alignment for nucleotide sequences". In: *Bioinformatics* 34.18 (2018), pp. 3094–3100. DOI: 10.1093/bioinformatics/bty191.

[97] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature methods* 9.4 (2012), pp. 357–359. DOI: 10.1038/nmeth.1923.

[98] Andrzej Zielezinski et al. "Alignment-free sequence comparison: benefits, applications, and tools". In: *Genome Biology* 18 (1 Dec. 2017), p. 186. DOI: 10.1186/s13059-017-1319-7.

[99] Siobain Duffy, Laura A Shackelton, and Edward C Holmes. "Rates of evolutionary change in viruses: patterns and determinants". In: *Nature Reviews Genetics* 9.4 (2008), pp. 267–276. DOI: 10.1038/nrg2323.

[100] Emidio Capriotti and Marc A Marti-Renom. "Quantifying the relationship between sequence and three-dimensional structure conservation in RNA". In: *BMC Bioinformatics* 11 (1 Dec. 2010), p. 322. DOI: 10.1186/1471-2105-11-322.

[101] Apurva Narechania et al. "Phylogenetic incongruence among oncogenic genital alpha human papillomaviruses". In: *Journal of virology* 79.24 (2005), pp. 15503–15510. DOI: 10.1128/JVI.79.24.15503-15510.20.

[102] Peter Norberg et al. "The IncP-1 plasmid backbone adapts to different host bacterial species and evolves through homologous recombination". In: *Nature Communications* 2 (1 Apr. 2011), p. 268. DOI: 10.1038/ncomms1267.

[103] Noboru Sueoka. "Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein". In: *Proceedings of the National Academy of Sciences* 47.8 (1961), pp. 1141–1149. DOI: 10.1073/pnas.47.8.1141.

[104] S Karlin, J Mrázek, and A M Campbell. "Compositional biases of bacterial genomes and evolutionary implications". In: *Journal of Bacteriology* 179 (12 June 1997), pp. 3899–3913. DOI: 10.1128/jb.179.12.3899-3913.1997.

[105] Rickard Sandberg et al. "Capturing Whole-Genome Characteristics in Short Sequences Using a Naïve Bayesian Classifier". In: *Genome Research* 11 (8 Aug. 2001), pp. 1404–1409. DOI: 10.1101/gr.186401.

[106] Rickard Sandberg et al. "Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content". In: *Gene* 311 (1-2 June 2003), pp. 35–42. DOI: 10.1016/S0378-1119(03)00581-X.

[107] Esko Ukkonen. "Approximate string-matching with q-grams and maximal matches". In: *Theoretical Computer Science* 92 (1 Jan. 1992), pp. 191–211. DOI: 10.1016/0304-3975(92)90143-4.

[108] Derrick E. Wood, Jennifer Lu, and Ben Langmead. "Improved metagenomic analysis with Kraken 2". In: *Genome Biology* 20 (1 2019), pp. 1–13. DOI: 10.1186/s13059-019-1891-0.

[109] Brian D Ondov et al. "Mash: fast genome and metagenome distance estimation using MinHash". In: *Genome Biology* 17 (1 Dec. 2016), p. 132. DOI: 10.1186/s13059-016-0997-x.

[110] Shahab Sarmashghi et al. "Skmer: Assembly-free and alignment-free sample identification using genome skims". In: *Genome Biology* 20 (1 2019), pp. 1–20. DOI: 10.1186/s13059-019-1632-4.

[111] David C Torney et al. "Computation of d_2: a measure of sequence dissimilarity". In: *Computers and DNA: The Proceedings of the Interface between Computation Science and Nucleic Acid Sequencing Workshop*. Routledge, 1990, pp. 109–125. DOI: 10.4324/9780429501463-11.

[112] Gesine Reinert et al. "Alignment-free sequence comparison (I): Statistics and power". In: *Journal of Computational Biology* 16 (12 2009), pp. 1615–1634. DOI: 10.1089/cmb.2009.0198.

[113] Nathan A Ahlgren et al. "Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences". In: *Nucleic Acids Research* 45 (1 Jan. 2017), pp. 39–53. DOI: 10.1093/nar/gkw1002.

[114] J. Qi, H. Luo, and B. Hao. "CVTree: a phylogenetic tree reconstruction tool based on whole genomes". In: *Nucleic Acids Research* 32 (Web Server July 2004), W45–W47. DOI: 10.1093/nar/gkh362.

[115] Qiang Li, Zhao Xu, and Bailin Hao. "Composition vector approach to whole-genome-based prokaryotic phylogeny: success and foundations". In: *Journal of Biotechnology* 149.3 (2010), pp. 115–119. DOI: 10.1016/j.jbiotec.2009.12.015.

[116] Leelavati Narlikar et al. "One size does not fit all: On how Markov model order dictates performance of genomic sequence analyses". In: *Nucleic Acids Research* 41 (3 Feb. 2013), pp. 1416–1424. DOI: 10.1093/nar/gks1285.

[117] Kujin Tang, Jie Ren, and Fengzhu Sun. "Afann: bias adjustment for alignment-free sequence comparison based on sequencing data using neural network regression". In: *Genome Biology* 20 (1 Dec. 2019), p. 266. DOI: 10.1186/s13059-019-1872-3.

[118] Peter Bühlmann and Abraham J. Wyner. "Variable length Markov chains". In: *The Annals of Statistics* 27 (2 Apr. 1999), pp. 480–513. DOI: 10.1214/aos/1018031204.

[119] Dana Ron, Yoram Singer, and Naftali Tishby. "The power of amnesia: Learning probabilistic automata with variable memory length". In: *Machine Learning* 25 (2-3 1996), pp. 117–149. DOI: 10.1007/BF00114008.

[120] Jorma Rissanen. "A universal data compression system". In: *IEEE Transactions on information theory* 29.5 (1983), pp. 656–664. DOI: 10.1109/TIT.1983.1056741.

[121] Dana Ron, Yoram Singer, and Naftali Tishby. "The power of amnesia: Learning probabilistic automata with variable memory length". In: *Machine learning* 25.2 (1996), pp. 117–149. DOI: 10.1023/A:1026490906255.

[122] Daniel Dalevi, Devdatt Dubhashi, and Malte Hermansson. "A new order estimator for fixed and variable length Markov models with applications to DNA sequence similarity". In: *Statistical Applications in Genetics & Molecular Biology* 5.1 (2006). DOI: 10.2202/1544-6115.1214.

[123] Alberto Apostolico and Gill Bejerano. "Optimal amnesic probabilistic automata or how to learn and classify proteins in linear time and space". In: *Journal of Computational Biology* 7.3-4 (2000), pp. 381–393. DOI: 10.1145/332306.332321.

[124] Gill Bejerano and Golan Yona. "Modeling protein families using probabilistic suffix trees". In: *Proceedings of the third annual international conference on Computational molecular biology*. 1999, pp. 15–24. DOI: 10.1145/299432.299445.

[125] Shaokun An et al. "A New Context Tree Inference Algorithm for Variable Length Markov Chain Model with Applications to Biological Sequence Analyses". In: *Journal of Computational Biology* 29.8 (2022), pp. 839–856. DOI: 10.1089/cmb.2021.0604.

[126] Peter Bühlmann. "Model selection for variable length Markov chains and tuning the context algorithm". In: *Annals of the Institute of Statistical Mathematics* 52.2 (2000), pp. 287–315. DOI: 10.1023/A:1004165822461.

[127] Daniel Dalevi, Devdatt Dubhashi, and Malte Hermansson. "Bayesian classifiers for detecting HGT using fixed and variable order markov models of genomic signatures". In: *Bioinformatics* 22 (5 Mar. 2006), pp. 517–522. DOI: 10.1093/bioinformatics/btk029.

[128] Mufleh Al-Shatnawi, M. Omair Ahmad, and M. N.S. Swamy. "Prediction of Indel flanking regions in protein sequences using a variable-order Markov model". In: *Bioinformatics* 31 (1 Jan. 2015), pp. 40–47. DOI: 10.1093/bioinformatics/btu556.

[129] Weinan Liao et al. "Alignment-free transcriptomic and metatranscriptomic comparison using sequencing signatures with variable length Markov chains". In: *Scientific reports* 6.1 (2016), p. 37243. DOI: 10.1038/srep37243.

[130] Matteo Comin and Morris Antonello. "On the comparison of regulatory sequences with multiple resolution Entropic Profiles". In: *BMC Bioinformatics* 17 (1 2016). DOI: 10.1186/s12859-016-0980-2.

[131] John A. Lees et al. "Fast and flexible bacterial genomic epidemiology with PopPUNK". In: *Genome Research* 29 (2 Feb. 2019), pp. 304–316. DOI: 10.1101/gr.241455.118.

[132] G Fichant and Christian Gautier. "Statistical method for predicting protein coding regions in nucleic acid sequences". In: *Bioinformatics* 3.4 (1987), pp. 287–295. DOI: 10.1093/bioinformatics/3.4.287.

[133] Diego Simón, Juan Cristina, and Héctor Musto. "Nucleotide Composition and Codon Usage Across Viruses and Their Respective Hosts". In: *Frontiers in Microbiology* 12 (June 2021). DOI: 10.3389/fmicb.2021.646300.

[134] Nina Mossadegh et al. "Codon optimization of the human papillomavirus 11 (HPV 11) L1 gene leads to increased gene expression and formation of virus-like particles in mammalian epithelial cells". In: *Virology* 326 (1 Aug. 2004), pp. 57–66. DOI: 10.1016/j.virol.2004.04.050.

[135] James T. Van Leuven et al. "ΦX174 Attenuation by Whole-Genome Codon Deoptimization". In: *Genome biology and evolution* 13 (2 2021), pp. 1–17. DOI: 10.1093/gbe/evaa214.

[136] Dusan Kunec and Nikolaus Osterrieder. "Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias". In: *Cell Reports* 14 (1 Jan. 2016), pp. 55–67. DOI: 10.1016/j.celrep.2015.12.011.

[137] Fiona Tulloch et al. "RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies". In: *elife* 3 (2014), e04531. DOI: 10.7554/eLife.04531.

[138] Valerie Odon et al. "The role of ZAP and OAS3/RNAseL pathways in the attenuation of an RNA virus with elevated frequencies of CpG and UpA dinucleotides". In: *Nucleic Acids Research* 47.15 (2019), pp. 8061–8083. DOI: 10.1093/nar/gkz581.

[139] Peter Simmonds et al. "Attenuation of dengue (and other RNA viruses) with codon pair recoding can be explained by increased CpG/UpA dinucleotide frequencies". In: *Proceedings of the National Academy of Sciences* 112.28 (2015), E3633–E3634. DOI: 10.1073/pnas.1507339112.

[140] Francesca Di Giallonardo et al. "Dinucleotide Composition in Animal RNA Viruses Is Shaped More by Virus Family than by Host Species". In: *Journal of Virology* 91 (8 Apr. 2017). Ed. by Terence S. Dermody. DOI: 10.1128/JVI.02381-16.

[141] Ariane Bize et al. "Exploring short k-mer profiles in cells and mobile elements from Archaea highlights the major influence of both the ecological niche and evolutionary history". In: *BMC Genomics* 22 (1 Mar. 2021), p. 186. DOI: 10.1186/s12864-021-07471-y.

[142] Clovis Galiez et al. "WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs". In: *Bioinformatics* 33 (19 Oct. 2017), pp. 3113–3114. DOI: 10.1093/bioinformatics/btx383.

[143] Shohei Kojima et al. "Virus-like insertions with sequence signatures similar to those of endogenous nonretroviral RNA viruses in the human genome". In: *Proceedings of the National Academy of Sciences* 118 (5 Feb. 2021), e2010758118. DOI: 10.1073/pnas.2010758118.

[144] Marcel H. Schulz et al. "Fast and Adaptive Variable Order Markov Chain Construction". In: vol. 5251 LNBI. Springer Berlin Heidelberg, 2008, pp. 306–317. DOI: 10.1007/978-3-540-87361-7_26.

[145] Marcel H. Schulz. "Personal communication". In: 2019.

[146] Gideon Schwarz. "Estimating the dimension of a model". In: *Annals of statistics* 6.2 (1978), pp. 461–464.

[147] Henry B Mann and Donald R Whitney. "On a test of whether one of two random variables is stochastically larger than the other". In: *The annals of mathematical statistics* (1947), pp. 50–60.

[148] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[149] Robert Giegerich, Stefan Kurtz, and Jens Stoye. "Efficient implementation of lazy suffix trees". In: *Software: Practice and Experience* 33.11 (2003), pp. 1035–1049. DOI: 10.1002/spe.535.

[150] Fabio Cunial, Jarno Alanko, and Djamal Belazzougui. "A framework for space-efficient variable-order Markov models". In: *Bioinformatics* 35 (22 Nov. 2019). Ed. by Inanc Birol, pp. 4607–4616. DOI: 10.1093/bioinformatics/btz268.

[151] Gene M Amdahl. "Validity of the single processor approach to achieving large scale computing capabilities". In: *Proceedings of the April 18-20, 1967, spring joint computer conference*. 1967, pp. 483–485. DOI: 10.1145/1465482.1465560.

[152] Marek Kokot, Maciej Długosz, and Sebastian Deorowicz. "KMC 3: counting and manipulating k-mer statistics". In: *Bioinformatics* 33.17 (2017), pp. 2759–2761. DOI: 10.1093/bioinformatics/btx304.

[153] Rajat Shuvro Roy, Debashish Bhattacharya, and Alexander Schliep. "Turtle: Identifying frequent k-mers with cache-efficient algorithms". In: *Bioinformatics* 30.14 (2014), pp. 1950–1957. DOI: 10.1093/bioinformatics/btu132.

[154] Guillaume Marçais and Carl Kingsford. "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers". In: *Bioinformatics* 27.6 (2011), pp. 764–770. DOI: 10.1093/bioinformatics/btr011.

[155] Morag C Timbury and JH Subak-Sharpe. "Genetic interactions between temperature-sensitive mutants of types 1 and 2 herpes simplex viruses". In: *Journal of General Virology* 18.3 (1973), pp. 347–357. DOI: 10.1099/0022-1317-18-3-347.

[156] Amanda M. Casto et al. "Large, stable, contemporary interspecies recombination events in circulating human herpes simplex viruses". In: *Journal of Infectious Diseases* 221 (8 2020), pp. 1271–1279. DOI: 10.1093/infdis/jiz199.

[157] David M. Koelle et al. "Worldwide circulation of HSV-2×HSV-1 recombinant strains". In: *Scientific Reports* 7 (1 Mar. 2017), p. 44084. DOI: 10.1038/srep44084.

[158] Axel Meyer et al. "Giant lungfish genome elucidates the conquest of land by vertebrates". In: *Nature* 590.7845 (2021), pp. 284–289. DOI: 10.1038/s41586-021-03198-8.

[159] B-H Juang and Lawrence R Rabiner. "A probabilistic distance measure for hidden Markov models". In: *AT&T technical journal* 64.2 (1985), pp. 391–408.

[160] Antonio Blanca et al. "The Statistics of *k*-mers from a Sequence Undergoing a Simple Mutation Process Without Spurious Matches". In: *Journal of Computational Biology* 29 (2 Feb. 2022), pp. 155–168. DOI: 10.1089/cmb.2021.0431.

[161] Hirotogu Akaike. "Information theory and an extension of the maximum likelihood principle". In: *Selected papers of hirotugu akaike* (1998), pp. 199–213.

[162] Nariaki Sugiura. "Further analysis of the data by Akaike's information criterion and the finite corrections: Further analysis of the data by akaike's". In: *Communications in Statistics-theory and Methods* 7.1 (1978), pp. 13–26.

[163] Jason Grealey et al. "The carbon footprint of bioinformatics". In: *Molecular biology and evolution* 39.3 (2022), msac034. DOI: 10.1093/molbev/msac034.

[164] Loïc Lannelongue, Jason Grealey, and Michael Inouye. "Green algorithms: quantifying the carbon footprint of computation". In: *Advanced science* 8.12 (2021), p. 2100707. DOI: 10.1002/advs.202100707.

[165] Kijong Yi et al. "Mutational spectrum of SARS-CoV-2 during the global pandemic". In: *Experimental & Molecular Medicine* 53 (8 Aug. 2021), pp. 1229–1237. DOI: 10.1038/s12276-021-00658-z.

[166] Lauriane de Fabritus et al. "Attenuation of tick-borne encephalitis virus using large-scale random codon re-encoding". In: *PLoS pathogens* 11.3 (2015), e1004738. DOI: 10.1371/journal.ppat.1004738.

[167] Megan H Powdrill et al. "Contribution of a mutational bias in hepatitis C virus replication to the genetic barrier in the development of drug resistance". In: *Proceedings of the National Academy of Sciences* 108.51 (2011), pp. 20509–20513. DOI: 10.1073/pnas.1105797108.

[168] Paul M Sharp et al. "Variation in the strength of selected codon usage bias among bacteria". In: *Nucleic acids research* 33.4 (2005), pp. 1141–1153. DOI: 10.1093/nar/gki242.

[169] Zane A Jaafar and Jeffrey S Kieft. "Viral RNA structure-based strategies to manipulate translation". In: *Nature Reviews Microbiology* 17.2 (2019), pp. 110–123. DOI: 10.1038/s41579-018-0117-x.

[170] Francisco P. Lobo et al. "Virus-Host Coevolution: Common Patterns of Nucleotide Motif Usage in Flaviviridae and Their Hosts". In: *PLoS ONE* 4 (7 July 2009). Ed. by Lark L. Coffey, e6282. DOI: 10.1371/journal.pone.0006282.

[171] F. García-Arenal and A. Fraile. "Trade-offs in host range evolution of plant viruses". In: *Plant Pathology* 62 (S1 Dec. 2013), pp. 2–9. DOI: 10.1111/ppa.12104.

[172] Amr El-Sayed and Mohamed Kamel. "Future threat from the past". In: *Environmental Science and Pollution Research* 28 (2021), pp. 1287–1291. DOI: 10.1007/s11356-020-11234-9.

[173] Jolyon M Medlock and Steve A Leach. "Effect of climate change on vector-borne disease risk in the UK". In: *The Lancet Infectious Diseases* 15.6 (2015), pp. 721–730. DOI: 10.1016/S1473-3099(15)70091-5.