Bioinformatics Advance Access published March 3, 2014

Challenges in RNA Virus Bioinformatics

Manja Marz[1,*], Niko Beerenwinkel[2,3], Christian Drosten[4], Markus Fricke[1], Dmitrij Frishman[5,6,7], Ivo L. Hofacker[8], Dieter Hoffmann[9], Martin Middendorf[10] Thomas Rattei[11], Peter F. Stadler[8,12,13,14], Armin Töpfer[2,3]

[1]Friedrich-Schiller-University Jena, Faculty of Mathematics und Computer Science, Leutragraben 1, 07743 Jena {manja, markus.fricke2}@uni-jena.de;
[2]Department of Biosystems Science and Engineering, ETH Zurich, Mattenstrasse 26, CH-4058 Basel, Switzerland, {niko.beerenwinkel,armin.toepfer}@bsse.ethz.ch;
[3]SIB Swiss Institute of Bioinformatics, CH-4058 Basel, Switzerland.
[4]Universitätsklinikum Bonn, Institut für Virologie, Sigmund-Freud-Str. 25, 53127 Bonn, drosten@virology-bonn.de;
[5]Technische Universität Muenchen, Wissenschaftszentrum Weihenstephan, Am Forum 1, 85354 Freising, Germany, d.frishman@wzw.tum.de
[6]Helmholtz Center Munich – German Research Center for Environmental Health (GmbH), Institute of Bioinformatics and Systems Biology, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany
[7] Moscow Institute of Physics and Technology, Institutskii Per. 9, Moscow Region, Dolgoprudny 141700, Russia [8]Institut für theoretische Chemie, Universität Wien, Währingerstraße 17, A-1090 Wien, Austria, ivo@tbi.univie.ac.at;
[9]Institute of Virology, Technical University of Munich, Munich, Germany, dieter.hoffmann@tum.de;
[10]University of Leipzig, Faculty of Mathematics and Computer Science, Institute for Computer Science, Augustusplatz 10, 04109 Leipzig
martin.middendorf@informatik.uni-leipzig.de;
[11]Division of Computational Systems Biology, Department of Microbiology and Ecosystem Science, Universität Wien, Althanstraße 14, 1090 Wien, Austria, thomas.rattei@univie.ac.at;
[12]Department of Computer Science and Interdisciplinary Center of Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany studla@bioinf.uni-leipzig.de;
[13]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany;
[14]The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe NM 87501, U.S.A.
* Corresponding author.

Associate Editor: Dr. Jonathan Wren

ABSTRACT

Computer-assisted studies of structure, function, and evolution of viruses remains a neglected area of research. The attention of bioinformaticians to this interesting and challenging field is far from commensurate with its medical and biotechnological importance. It is very telling that out of over 200 talks held at ISMB 2013, the largest international bioinformatics conference, only one presentation explicitly dealt with viruses. In contrast to many broad, established and well organized bioinformatics communities (e.g. structural genomics, ontologies, next-generation sequencing, expression analysis), research groups focusing on viruses can probably be counted on the fingers of two hands. The purpose of this review is to increase awareness among bioinformatics researchers about the pressing needs and unsolved problems of computational virology. We focus primarily on RNA viruses that pose problems to many standard bioinformatics analyses due to their compact genome organization, fast mutation rate, and low evolutionary conservation. We provide an overview of tools and algorithms for handling viral sequencing data, detecting functionally important RNA structures, classifying viral proteins into families, and investigating the origin and evolution of viruses.

1 INTRODUCTION

Viruses have been the first biological systems for which complete genomic information became available: bacteriophage MS2 in 1976, and ΦX174 in 1977. By the mid 1990s multiple strains of important human pathogenes, in particular HIV, had been sequenced, laying the foundation for a systematic comparative genomics of the diverse virus families. With the completion of the human genome and the sequencing of 100s of eukaryotic and 1000s of prokaryotic genomes, however, the bioinformatics community focussed almost

entirely on these much larger and more complex systems. With few exceptions, most viral genomes are thus relatively poorly annotated and few compational tools and techniques have been developed specifically for the many idiosyncratic features of individual virus families.

The small size of viral genomes makes it possible to sequence large numbers of isolates, usually in clinical context, that are unavailable for any living systems. This flood of sequencing data in itself calls for specific methods of analysis, which so far are available in part at best.

In this survey we concentrate on RNA viruses. They form a highly diverse grouping, usually classified in terms of their genome organization. They may have a single stranded (ss) genome in either plus (e.g. Poliovirus, Enterovirus, Hepatitis-A-virus), or in minus orientation (e.g. Rabiesvirus, maize-mosaic-virus), or a double-stranded (ds) RNA genome (e.g. Rotavirus), or ssRNA with a double stranded (ds) DNA as an intermediate product (e.g. Retroviruses including HIV). Nevertheless, they share several common features. In particular, their genomes are small, ranging from 3,400 nt (Enterobacteria phage BZ13) to 31.000 nt (coronavirus), encoding just a a handful of proteins. Furthermore, their mutation rates are large enough to form a quasispecies rather than a single, genetically homogenous species in the classical sense.

Many of the common questions raised in virology call for specific bioinformatics support: How is viral gene expression regulated? How do RNA viruses evolve? How common are they? Have we already seen the whole diversity of RNA virus families or genera? How quickly do they change? How often do variations occur? How important is recombination in viral evolution? Is there a single common viral origin or do we find clearly independent origins?

In this contribution we review the state of the art of the computational methods that have helped to address some of these questions. In particular, we will be concerned with (1) finding and assembling viral genomes based on RNA-seq data, (2) regulatory RNA elements and processed subgenomic RNA species, (3) classification problems regarding viral protein families, (4) phylogenetics and evolution of RNA viruses, and (5) virus-host associations as a basis for biomedical applications.

2 DISCOVERY OF VIRAL SEQUENCES

Viruses display high genetic diversity both within and among viral species as well as within and among infected hosts. Although next-generation sequencing provides cost effective access to high throughput data, inferring the viral genetic diversity of a mixed sample from deep-coverage sequencing data has remained a challenging task. The reasons lie in difficulties of sample preparation and sequencing errors, short read length, and in particular a very incomplete *a priori* knowledge of existing viruses and their diversity. Viral diversity estimation may range from identifying viral species in metagenomics studies to reconstructing the individual mutants in the intra-host population of a single species.

The composition of a mixed sample can be assessed by metagenomics approaches, reviewed in (Fancello *et al.*, 2012), (Mokili *et al.*, 2012), and (Reyes *et al.*, 2012). A main approach is sequence read annotation by taxonomic classification using existing reference genomes and databases [1]. In many metagenomics applications, however, classification is not possible, because the majority of sequences have no known reference genome or homolog (Edwards & Rohwer, 2005). In this case, *de novo* discovery of viral species can be performed by state-of-the-art *de novo* genome assemblers. These methods try to assemble the genomes of the major species in the sample, ignoring low-frequency variants and technical errors.

Once the (reference) genomes are known and reads are classified, the resolution of diversity estimation can be increased by inferring the viral population structure of each individual species. Intra-host virus populations consist of many related mutants, generated by mutation, recombination, and selection. Even low-frequency variants can be of great interest, for example, because they may harbor drug resistance mutations (Barzon *et al.*, 2011), facilitate immune escape (Luciani *et al.*, 2012), or affect virulence (Töpfer *et al.*, 2013*a*). Estimating intra-host viral genetic diversity and reconstructing the individual haplotype sequences relies on both error correction and read assembly. It can be performed on different spatial scales, including single sites of the genome (Single-Nucleotid-Varient calling), small sliding windows (local reconstruction), or complete genomes (global reconstruction). Current viral haplotype reconstruction tools, reviewed in (Beerenwinkel *et al.*, 2012), (Beerenwinkel & Zagordi, 2011), and (Vrancken *et al.*, 2010), can quantify viral diversity from NGS data, with recombinant population structure (Töpfer *et al.*, 2013*b*), provided that haplotypes differ enough, reads are not too short, and coverage is high(Zagordi *et al.*, 2012). A common prerequisite for these tools is a high-quality alignment of the reads.

As of today, NGS-based discovery of viral sequences in mixed samples remains challenging, because most analysis steps are not easily automated and each one has technical or biological limitations. There is a need for an integrated workflow combining the different processing steps in viral diversity studies to discover the underlying virus populations that can be used on a daily basis by clinicians and virologists.

3 STRUCTURAL RNA ELEMENTS

3.1 Detection and Distribution of Structured RNAs

The realization that conserved RNA structure plays a role in virology dates back to the beginning of the 1980s (Ahlquist *et al.*, 1981). Most of the structured viral RNA elements contained in the `Rfam` database are cis-acting elements, in particular internal ribosomal entry sites (IRES), cis-acting replication elements (CRE), and other elements located in the untranslated regions (UTRs) of RNA viruses. Functional RNA structures also appear to be abundant within the viral coding regions. Furthermore, regular arrangements of hairpins throughout the genomic RNA have been shown to be instrumental for packaging in Leviviridae (Dykeman *et al.*, 2011) and some satellite viruses (Schroeder *et al.*, 2011). Evolutionary conserved large-scale ordering of RNA virus genomes seems to be abundant in many animal and plant viruses (Davis *et al.*, 2008).

The first systematic searches for conserved, and hence likely functional, RNA secondary structure elements were performed

[1] http://www.rna.uni-jena.de/rna.php

in RNA viruses more than a decade ago (Rauscher *et al.*, 1997). This stimulated the development of early computational methods (Hofacker *et al.*, 1998; Hofacker & Stadler, 1999) capable of surveying alignments of complete virus genomes (Witwer *et al.*, 2001; Thurner *et al.*, 2004) for local RNA motifs in which the structure is more conserved than the underlying sequence. Somewhat surprisingly, however, the next generation of comparative RNA secondary structure predictors such as `RNAz` (Washietl *et al.*, 2005) and `evofold` (Pedersen *et al.*, 2006) apparently have not been used extensively on virus data. The results of (Davis *et al.*, 2008) suggest that coverage with conserved secondary structure varies substantially between virus families.

Viral RNAs have recently become accessible to structural probing at larger scales using combinations of SHAPE and sequencing. The analysis of these data requires both elaborate processing of the raw SHAPE data (Pang *et al.*, 2011), and the incorporation of these data into RNA structure prediction algorithms in the form of constraints (Reuter & Mathews, 2010; Washietl *et al.*, 2012). First results include the HCV 5' UTR (Pang *et al.*, 2011) and the secondary structure of a complete HIV-1 genome (Watts *et al.*, 2009). Since essentially all RNA molecules form secondary structures, one has to keep in mind that the entire structure is not necessarily of functional relevance.

3.2 Viral ncRNAs

In addition to proteins, viruses may also encode non-coding RNAs (ncRNAs). Although most of the well-described examples have been found in viruses with DNA genomes, we include a brief overview here for two reasons: First, ncRNAs do appear in retroviruses and second it is at least conceivable that processing products of viral RNAs might act as ncRNA species.

The formation of independent, functional RNA species is a widespread phenomenon among diverse virus families, best known but apparently not limited to DNA viruses (Table 1). The largest class are virus-encoded microRNAs, of which more than 200 distinct types have been reported over the past decade (Grundhoff & Sullivan, 2011; Grundhoff, 2011), with herpesviruses accounting for the overwhelming majority of examples (Boss *et al.*, 2009). Smaller numbers of examples have been reported also in Polyomaviridae, Ascoviridae, Baculoviridae, and Retroviridae. Viral microRNAs appear to regulate viral-encoded transcripts and/or networks of host genes predominantly using the host miRNA regulation systems.

In contrast to animal and plant microRNAs, their viral counterparts are often poorly conserved. This complicates both their annotation in newly sequenced genomes (Grundhoff & Sullivan, 2011; Grundhoff, 2011) and the computational reconstruction of their interaction networks (Ghosh *et al.*, 2009; Kim *et al.*, 2012). Nowadays, new viral microRNAs are usually found by means of deep sequencing, see e.g. (Tuddenham *et al.*, 2012). A recurring computational problem in this context is to distinguish *bona fide* microRNAs from small degradation products.

The RNA repertoire of herpesvirus species is by no means restricted to microRNAs, Tab. 1. They also encode a diverse set of small nuclear RNAs with diverse functions, a small nucleolar RNA, and even a derived copy of the telomerase RNA component. In some cases unrelated ncRNAs from different families have analogous functions. For instance, both the EBER-1 RNA of the herpesvirus EBV and the VA-I RNAs of adenoviruses effective inhibitors of PKR activation (McKenna *et al.*, 2006). Again, the oftentimes poor conservation and the diversity of the viral RNAs complicats their annotation.

Table 1. Examples of virus-encoded small RNAs.

ncRNA	Virus	Reference
MicroRNAs		
BART cluster	Epstein-Barr	Edwards *et al.* (2008)
BHRF1 cluster	Epstein-Barr	Pfeffer *et al.* (2004)
TAR-mir	HIV-1	Klase *et al.* (2007)
Small nuclear RNAs		
EBER-1,2	Epstein-Barr v.	Klase *et al.* (2007)
HSURs 1-6	Herpes saimiri v.	Klase *et al.* (2007)
VA-I,II	Adenoviruses	Mathews (1995)
telomerase RNA	Marek's disease v.	Fragnet *et al.* (2005)
v-snoRNA-1	Epstein-Barr v.	Hutzinger *et al.* (2009)
Long ncRNAs		
PAN	KSAH	Klase *et al.* (2007)

A related topic are subviral RNAs and satellite RNAs. In particular plant viruses often bring with them non-coding deletion mutants. These defective interfering (DI) RNAs often maintain crucial cis-acting RNA elements (Pathak & Nagy, 2009), which are described in more detail below.

3.3 Secondary structures in the mRNA coding regions

The existence of extensive secondary structures in native mRNAs is well supported by experimental evidence, and *in silico* with the assumption that they have lower folding energies and are thus more stable than codon-randomized sequences (Katz & Burge, 2003). However, in general and especially for comparatively variable viral sequences MFE is considered to be minor relevant (Rivas & Eddy, 2000; Workman & Krogh, 1999) and compensatory mutations analysis over a broader range of individuals would be more sophisticated. On the other hand, computational analysis suggests that the three mRNA functional domains – 5'UTR, CDS, and 3'UTR – form largely independent folding units while base pairing across domain borders is rare (Shabalina *et al.*, 2006). Global architectures appear to be poorly conserved between sequence-similar mRNA molecules (Chursov *et al.*, 2012*b*), but evolutionary conserved functional local secondary structures are abundant (Meyer & Miklós, 2005; Olivier *et al.*, 2005; Findeiss *et al.*, 2011). The relationship between mRNA structure and gene expression has been demonstrated both computationally and experimentally (Kudla *et al.*, 2009; Duan *et al.*, 2003; Ilyinskii *et al.*, 2009; Carlini *et al.*, 2001; Nackley *et al.*, 2006). For example, in the influenza virus a novel structural feature was identified in a functionally important region of the NS1 mRNA (Ilyinskii *et al.*, 2009). Synonymous mutations altering this mRNA element lead to significantly reduced protein expression while non-synonymous mutations designed to preserve this local structure do not affect expression, implying that distinct secondary structure elements may be important for viral gene expression. Reduced mRNA stability near the start codon has been observed in a wide range of species, including dsDNA viruses (Zhou & Wilke, 2011), probably as a mechanism to facilitate ribosome binding or start codon recognition

by initiator-tRNA (Gu *et al.*, 2010). There is also computational evidence that temperature-induced changes in mRNA structures may constitute a yet unappreciated molecular mechanism of the virus cold adaptation/temperature sensitivity phenomena (Chursov *et al.*, 2012*a*).

In a few cases, extensive and well-conserved RNA structures are superimposed on the coding sequence. Maybe the most impressive example is the IRES of HIV-2 (Herbreteau *et al.*, 2005) and the Rev-response element (RRE) in HIV-1 (Pallesen *et al.*, 2009). Internal RNA elements are also located in the ORF1b of group 2 coronavirus MHV; here deletion analysis has identified a 69-nt bulged stem loop required for packaging RNAs into particles (Fosmire *et al.*, 1992) or the cis-active elements involved in picornavirus replication (Steil & Barton, 2009). The latter initiate plus and minus strand RNA synthesis and cloverleaf elements, controlling both translation and replication (Liu *et al.*, 2009*a*). Another example is the ribosomal frameshift known e.g. in coronavirus ORF1, induced by a short hairpin of 4–11 basepairs, which also affects genomic and subgenomic RNA production (Plant *et al.*, 2013).

One particularly intriguing aspect of mRNA life – the one that makes it distinctly different from any other kind of RNA – is the dual selection pressure towards maintaining both stable RNA structures of CDSs and the three-dimensional folds of their encoded proteins (White *et al.*, 1972). Additional layers of selection arise due to microRNA and protein binding sites within the mRNA coding regions. It has been argued that the redundancy of the genetic code plays an important role in satisfying these requirements (Shabalina *et al.*, 2006). Evolutionary models able to make a distinction between the evolutionary pressure at the RNA and protein level have been proposed (Rubinstein *et al.*, 2011).

In general structure prediction of the mRNA coding regions remains an under-appreciated area of RNA bioinformatics, arguably because these molecules are large and do not easily yield to current structure prediction methods and experimental structure-probing data is only beginning to emerge (Kertesz *et al.*, 2010). Most of the insights into the evolutionary constraints acting on mRNAs therefore come from correlating predicted base-paring patterns with the effects of site-directed mutagenesis on mRNA expression and degradation as well as on the expression levels and activity of encoded protein products. On the other hand prediction of mRNA secondary structure is facilitated by the availability of abundant comparative sequence information both from viruses and cellular organisms. `RNAdecoder` (Pedersen *et al.*, 2004; Meyer & Miklós, 2005) implements a comparative method for finding and folding RNA secondary structures within protein-coding regions. A recent survey of fly genomes (Findeiss *et al.*, 2011), however, indicated that the specificity of this combined approach is insufficient for genome-wide application. The problem of cataloging conserved secondary structure motifs within coding regions, in particular in viruses, remains open.

3.4 The Secrets of Viral UTRs

Functionally important viral RNA structures tend to be concentrated in the UTRs. This is not unexpected, of course, since UTRs are typically the only non-coding regions within the densely packed virus genomes. A wide range of experimental data indeed demonstrates that the UTRs are essential for determining the efficiency of translation, mRNA life time, and localization. Functional UTR elements are often binding sites for viral or host proteins, but can also be involved in RNA-RNA interactions either within the genome (cyclization) or with host RNAs (e.g. the ribosome). UTR structures have been studied most intensively in positive-strand RNA viruses (Liu *et al.*, 2009*b*), in particular in those affecting humans and/or animal livestock.

In positive-strand RNA viruses the genomic RNA has to function directly as an mRNA. However, the viral RNA often lacks the 5'cap as well as the poly-A tail of canonical eukaryotic mRNAs. Eukaryotic translation usually starts with binding of initiation factors to the 5'cap, these in turn recruit the small ribosomal subunit, which then scans along the mRNA. In 1988 two independent studies (Pelletier & Sonenberg, 1988; Jang *et al.*, 1988) showed that certain picornaviruses exhibit a cap-independent translation initiation mechanism. The structured RNA region of some 300 nt to 700 nt responsible for this mechanism was termed "internal ribosome entry site" (IRES), and is perhaps the best studied example of a viral UTR structure. Indeed, IRES structures seem to be present in all *Picornaviridae* (Witwer *et al.*, 2001). While viruses in the genus Flavivirus have a 5'cap, the other genera of the family *Flaviviridae*, such as Pestivirus and Hepacivirus, seem to use IRES structures (Thurner *et al.*, 2004). In addition IRESs have been found or implicated in several other viruses including the subgenomic mRNAs of retroviruses, see e.g. (Vallejos *et al.*, 2012). While IRES regions within e.g. the *Picornaviridae* show significant similarity of RNA secondary structure, no such similarities are obvious across family boundaries, arguing against a common origin of different IRES structures.

IRES structures are also common in some positive strand RNA plant viruses, while others replace the IRES with a structure in the 3'UTR, called 3'cap-independent translation enhancers (3'CITE) (Nicholson & White, 2011). 3'CITE structures are much shorter (around 100nt) and can be grouped into several distinct classes. Some of them bind translation initiation factors such as elF4E, while others seem to interact directly with the ribosome. Cyclization of the RNA is required to then bring these initiation factors/ribosome close to the translation start site.

Genome cyclization through complementary sequences in the 5' and 3'UTR is a common theme observed in many virus families. The flaviviruses are an example among the positive strand RNA viruses. Presumably, cyclization improves translation rates by allowing the ribosome to transfer from the 3'end back to the start of the coding region. Among negative strand viruses, the *Bunyaviridae*, including Hanta virus, or *Orthomyxoviridae*, including Influenza virus exhibit a segmented genome where each segment has strong complementarity between 3' and 5' end. In UTRs of plant virus genomes or HCV tRNA-like secondary structures are known, which are believed to interact with the viral genome and the ribosome may interact during translation (Annamalai & Rao, 2006; Piron *et al.*, 2005).

The 5'UTR and the 3'UTR featuring translation efficiency and replication. Both *Picornaviridae* and *Flaviviridae* contain highly structured UTRs, which however differ significantly between genera. In Enterovirus these structures have been shown to be essential for the assembly of the RNA-dependent RNA polymerase (RdRp) complex (Zoll *et al.*, 2009).

3.5 Cis-acting elements

Apart from target prediction for viral microRNAs, interactions of structured RNA elements in viruses have remained largely unexplored. 5' and 3' UTRs containing cis-active elements are essential for viral genome replication (van den Born & Snijder, 2008; Ulferts & Ziebuhr, 2011). A complex example is given by nidoviruses, which synthesize a nested set of 3'/5'-coterminal subgenomic (sg)mRNAs (van Berlo *et al.*, 1982; Stern & Kennedy, 1980) from which the structural and accessory protein genes are expressed. The 5' ends of nidovirus sgRNAs share a leader sequence that is identical to the 5'-end of the genomic RNA (van Vliet *et al.*, 2002; de Vries *et al.*, 1990; Spaan *et al.*, 1983). A copy of this leader sequence is fused to the 3'-ends of nascent sg minus-strand RNAs in a process called discontinuous extension of minus strands (Sawicki & Sawicki, 1995; Sawicki *et al.*, 2007). The so-called transcription regulation sequences (TRSs) are located upstream of each of the structural and accessory protein genes, however, TRSs is also found downstream of the 5'-leader sequence on the viral genomic RNA (Ulferts & Ziebuhr, 2011). The proposed coronavirus transcription mechanism implies a close interaction between TRS-L and each of the cTRS-B present in the genomic RNA, imposing strong constraints of the evolution of the TRS sequences (Zúñiga *et al.*, 2004; Enjuanes *et al.*, 2001). The hypothesis of sgmRNA synthesis in coronaviruses requires a minimum thermodynamic stability in the TRS-L and cTRS-B duplex (Sola *et al.*, 2005; Dufour *et al.*, 2011) could not been proven *in silico* for all coronaviruses (Fricke & Marz, 2013). Corona-, Bafini- and Arteriviruses feature such a leader sequence, while Okaviruses do not (Cowley *et al.*, 2002); Equine torovirus contains one sgRNA with a 5'-leader sequence while the other sgmRNA species are leaderless (van Vliet *et al.*, 2002).

Coronavirus harbors several additional cis-active elements forming distinct stem loops involved in regulating sgRNA transcription and RNA replication (Li *et al.*, 2008; Raman *et al.*, 2003; Raman & Brian, 2005; Liu *et al.*, 2009*a*). The 3'UTR of mouse hepatitis virus (MHV) and BoCV upstream end contains a bulged stem loop and a pseudoknot, which cannot form simultaneously. This has led to a proposal that these structures are part of a molecular switch that regulates different steps of replication (Hsue & Masters, 1997; Williams *et al.*, 1999). Several cis-active structures are located within the coding regions.

The examples given here are by no means exhaustive. In fact the Rfam database features several dozens of distinct families. Despite their importance for viral control, however, there is no comprehensive analysis and only a few computational surveys, see e.g. (Li *et al.*, 2010), have been attempted following a few family-specific studies almost a decade ago (Witwer *et al.*, 2001; Thurner *et al.*, 2004; Hofacker *et al.*, 2004). Comparative investigations across families and detailed studies into the evolution of these elements are largely lacking.

4 CLASSIFICATION OF VIRAL PROTEIN FAMILIES

Already early in the era of genomics, NCBI's viral genomes project established a large-scale comparative resource providing information on orthology and paralogy of viral proteins (Bao *et al.*, 2004). Clusters of related viral proteins (viral orthologous groups, VOG), as well as the specialized collection of phage orthologous groups (POG) are available as part of the NCBI protein clusters (Sayers *et al.*, 2012; Kristensen *et al.*, 2013). Widely used databases of orthologous proteins such as EggNOG (Powell *et al.*, 2012), OMA (Altenhoff *et al.*, 2011) or KEGG (Kanehisa *et al.*, 2012), do not consider virus proteins at all. A variety of software tools for orthology detection have been proposed (Koonin, 2005), falling into four large groups (Kristensen *et al.*, 2010): phylogenetic tree-based approaches, heuristic best-match methods, synteny-based, and hybrid approaches. None of the available tools such as `EnsemblCompara` (Vilella *et al.*, 2009), `OrthoMCL` (Li *et al.*, 2003), or `InParanoid` (Alexeyenko *et al.*, 2006) have been specifically designed for viral genome analysis.

Remarkably, even the VOG data are rarely used for comparative genomics of viruses, despite their potential for studying the natural history of viral genes (Koonin *et al.*, 2006). So, what are the problems with the currently available VOG? Three main limitations are most evident: current VOG (a) are rarely updated and are not hierarchical, (b) lack remote and short homologs, and (c) are not linked to their hosts and other cellular organisms.

A significant fraction of current limitations in comparative genomics of viruses derives from the very divergent sequences. Analysis tools which include a broad analysis of compensatory mutations of a wide range and conserved motifs for further interactions do not exist and would results in higher computational costs of the underlying calculations. Recently developed tools such as `Phamerator` rely on fast, but even less sensitive approaches for sequence similarity calculations (Cresawn *et al.*, 2011). Considering the explosive growth of the genome databases, such all-*versus*-all comparisons can be expected to become even more crucial in the future. Efficient approaches such as incrementally calculated matrices of sequence similarities (Arnold *et al.*, 2005) are therefore promising tools for the next generation of classification systems for viral protein families.

5 VIRUS EVOLUTION AND PHYLOGENETICS

Phylogenetic analysis is a ubiquitous method in virology, forming an essential element of investigations describing viruses or viral epidemiology. However, several characteristics of viruses pose specific challenges for phylogenetics: i) strong differences in evolution rates, typically high on a short term qand much lower on the long term, ii) large potential for recombination and gene transfer even between distant viruses or their host species, iii) often strong evolutionary relationships between viruses and their hosts iv) lack of physical 'fossil records' of viruses, and v) abundance of genomic 'fossil records' (viral fossiles) as parts of ancient viral genomes that occur within the genomes of extant species.

Phylogenetic trees are the most wide-spread presentation for virus phylogenies in the literature and several tree-building methods and software exists (e.g., `MrBayes` (Ronquist & Huelsenbeck, 2003), `BEAST` (Drummond *et al.*, 2012), `PhyloBayes` (Lartillot *et al.*, 2009), `RAxML` (Stamatakis *et al.*, 2008)). However, trees cannot represent complex evolutionary relations that are relvent for viruses as horizontal gene transfer, interspecific recombination, or the evolutionary relations between viruses and their hosts. Different types of phylogenetic networks have been developed in recent years to represent such relations (e.g., (Huson *et al.*, 2011)). But there is

still much need for research on how to reconstruct such aspects of virus phylogeny.

5.1 Short-Term Viral Evolution

The short-term evolution rates of many viruses are so high that genomic evolution can already be observed over the course of years or even days. For analysing viral short-term sequence evolution it is important that the phylogenetic methods can include the sampling dates of the sequences (e.g., `TipDate` (Rambaut, 2000)). Moreover, spatial dispersal processes play an important role, e.g. the spatial distribution of a virus within the hosts body or the geographic spread of an infectious desease. Several methods and tools have been developed to analyse and reconstruct the history of such complex phylogenetic and phylogeographic processes (overviews are (Bloomquist *et al.*, 2010), (Faria *et al.*, 2011), (Lemey *et al.*, 2009)). Recent tools that implement Bayesian approaches are based on Markov chain models or continuous diffusion models (`BayArea` (Landis *et al.*, 2013), `SPREAD` (Bielejec *et al.*, 2011)). It is not easy to interprete the delivered phylogeographic reconstructions and visualization tools, e.g. `Phylowood` (Landis & Bedford, 2014), can help.

The evolutionary rates of virues can differ even for short-term evolutionary scenarios, for example, between different lineages (infectious, non-infectious) or between different time intervalls (e.g., states of an infection or seasons). One reason is that substitution rates reflect a complex product of mutation rate, generation time, effective population size and fitness (Sanjuán *et al.*, 2010; Jenkins *et al.*, 2002). Vastly different replication profiles (stamping machines *vs.* geometric replication (Martínez *et al.*, 2011)), make the estimation of substitution rates difficult. In viruses in particular, substitutions may also be an artifact caused by polymerase errors and nucleotide modifications (Domingo & Holland, 1997). For all these reasons, the classical assumption of a time-homogeneous substitution processe that is used by several phylogenetic and phylogeographic statistical inference methods does not hold and new approaches that can include varying evolutionary rates have been proposed, such as (Bielejec *et al.*, 2013). Unfortunately, the computational effort of such complex statistical inference methods is very high. One remedy is to use parallelized versions that could offer a dramatic speed up on various parallel architectures, e.g. computer cluster (Baele & Lemey, 2013), graphics processing units (GPUs) with `BEAGLE` (Ayres *et al.*, 2012), or multiple Field Programmable Gate Arrays (FPGAs) via `extended BEAGLE` (Jin & Bakos, 2013).

5.2 Viral "Deep Phylogeny"

Since physical fossiles of viruses do not exist there is no direct evidence about the time when viruses have emerged and their origin is still not clear. There is indication that viruses are polyphyletic and several hypothesis exists about their relation to cellular life (Wessner, 2010): i) they might be precursors of cellular life or ii) they might have originated from cellular life via a regressive, or reductive, process from whole cells or via a progressive process from genetic elements.

A problem for "deep phylogeny" reconstruction is that the genetic distance between viruses can be so large that reasonable alignments become impossible to calculate and therefore standard alignment based phylogenetic methods cannot be applied. The development of advanced approaches to achieve biologically correct alignments would help, but can only marginally alleviate the problem of saturated substitution processes. Other, approaches might be based on using aspects of genome organization or protein structures as phylogenetic characters (Holmes, 2011).

However, some ancient virures have left parts of their genome (or other traces) in the genome of germe line cells of their hosts. Such parts, called endogenous viral elements (EVEs), have survived as non-functional, neutrally evolving pseudogenes or even became fixed as functional. Most EVEs stem from retroviruses because they integrate into host genomes as part of their life cycle. For example about 8% percent of the human genome is derived from over 100,000 retroviral fossiles (Lander *et al.*, 2001). But in recent years also EVEs from many other viruses have been found (Horie & Tomonaga, 2011; Katzourakis & Gifford, 2010; Patel *et al.*, 2011). Some paleoviruses could even be almost entirely reconstructed from EVEs. To detect EVEs in complete genome sequences different programs have been developed, e.g., `RepeatMasker` (Smit *et al.*, 2010), `LTR_STRUC` (McCarthy & McDonald, 2003), `RetroTector` (Sperber *et al.*, 2009), and using a combination of several of them seems most promising (Lerat, 2009).

Orthologues EVEs that are found in multiple host species indicate a single integration event that happend before the divergence of the host species group and therefore can be used to infer the phylogeny of ancient viruses and to calibrate the long-term evolutionary timelines for viruses (Feschotte & Gilbert, 2012). With EVEs it was possible to stretch back the history of several RNA virus families (e.g. bornaviruses (Horie & Tomonaga, 2011)) over some 40 million years. This example shows that EVEs might help to solve the following problem of RNA virus phylogenetic dating. Studies that are based on genomic sequences of extant species often came to the conclusion that large taxonomic units of viruses (on the rank of genera) must have evolved from a common ancestor several ten-thousand years ago, whereas there are contrasting ideas based on virus-host coevolution over similarly wide taxonomic entities, suggesting bifurcation ages in the range of several millions of years (Buckling & Brockhurst, 2012; Fraile & García-Arenal, 2010; Marques & Carthew, 2007). However, in general, the calibration of phylogenies with fossile dates is difficult when the evolutionary rates are heterogenous and new algorithmic methods have to be developed for this (see (Heath *et al.*, 2013)).

5.3 Virus-Host Associations

Associations between viruses an their hosts can have an important influences on the phylogeny of both partners. A divergence of the host might lead to a divergence of the virus (codivergence) and hence to a (local) congruence of both phylogenies. Such a match of the virus phylogeny with host evolutionary events at known dates can be used to calibrate the virus phylogeny or corresponding molecular clocks (Sharp & Simmonds, 2011). The property of viruses to switch their hosts may enable viruses to replicate and spread much more efficiently, a process commonly referred to as an epidemic is observed in pathogenic viruses (Weiss, 2003). Due to the advantages conferred by the conquest of new host territory, some researchers assume that host switching is an elementary component of virus evolution and might also initiate viral speciation (Kitchen *et al.*, 2011).

Since virologists are highly interested to reconstruct the common history of viruses and their hosts several bioinformatics tools have been developed for this purpose (for an overview see (de Vienne *et al.*, 2013; Doyon *et al.*, 2011)). A program for testing of congruence between host phylogenies and parasite phylogenies is `ParaFit` (Legendre *et al.*, 2002). A fast implementation of `ParaFit` (`AxParafit` (Stamatakis *et al.*, 2007)) is integrated via a wrapper in the tool `CopyCat` (Meier-Kolthoff *et al.*, 2007), which incorporates also a graphical user interface. `CopyCat` was used, e.g. to investigate the codivergence between mycoviruses and their hosts (Göker *et al.*, 2011).

Most programms for inferring reconciliations use a parsimony criterion where a reconciliation of minimum total cost is sought for. In this approach a cost is given to each evolutionary event type (e.g., codivergence or host switch) and the total cost of a reconciliation is the sum of the costs of its events. The most often used programms are `CoRe-Pa` (Merkle *et al.*, 2010), `Jane` (Conow *et al.*, 2010) and `TreeMap` (Charleston & Page, 2002). An evaluation of the different reconciliation programs using a new model for cophylogeny generation can be found in (Keller-Schmidt *et al.*, 2011). These programs have been used, e.g., to reconstruct the phylogenetic relationship between orbiviruses (Dilcher *et al.*, 2012), papillomaviruses (Gottschling *et al.*, 2011), or arenaviruses in Africa (Coulibaly-N'Golo *et al.*, 2011) and their respective hosts.

There are still many research problems and a need for new bioinformatics methods that can, e.g. include biogeographic information and ecological traits, preferential host switching (Cuthill & Charleston, 2013) or different mutation rates. A better knowledge of the timing and underlying conditions of those processes could enable projections into the future and thereby contribute to the tackling of one of the major issues in today's infectious diseases research, i.e., the prediction (and prevention) of future pandemics and outbreaks.

6 MEDICAL AND BIOTECHNOLOGICAL APPLICATIONS

Viral evolution has many implications for clinical virology. Emergence of resistance mutations is among the biggest obstacles to a successful antiviral therapy (Richman, 2006). The molecular mechanisms selecting resistance mutations are complex, particularly when multiple antiviral agents are used, e.g. for HIV (Shibata *et al.*, 2011). Computational analysis of HIV-1 genome variation correlating with CCR5 or CXCR4 coreceptors led to AIDS therapy strategy (Lengauer *et al.*, 2007).

Thus new bioinformatics approaches to characterize viral evolution both on an intra- and interindividual level would be an important asset. A better understanding and knowledge of additional selection mechanisms such as RNA secondary structures could explain new pathways leading to resistance and immune escape mutations. New sequence based therapeutic concepts, such as RNA interference could prevent the selection of resistance mutations (Schopman *et al.*, 2012).

Previous work showed that the host immune system influences the genetic variability in chronically infected individuals (Hoffmann *et al.*, 2008). This can be analyzed with population genetic methods (Hoffmann *et al.*, 2012). In this setting viral species acquire numerous mutations over time (Hoffmann *et al.*, 2010). An increasing number of immunocompromized patients susceptible to chronic infections represents an important reservoir for new viral genotypes (Siebenga *et al.*, 2008).

By integrating bioinformatic methods it might be possible to predict viral evolution in patients from their individual viral population, including lower prevalent individuals with single genetic variations. Thus the goal is to forecast the course of a virus infection and adjust the treatment accordingly.

7 CONCLUDING REMARKS

RNA viruses pose a wide variety of challenges to computational methodology owing to their staggering diversity, compact genome organization, and rapid rate of evolution. On the other hand, the availability of large numbers of complete sequences and the small to at most moderate size of RNA virus genomes holds particular promises for specialized bioinformatics approaches. The latter two characteristics are shared with (animal) mitochondrial genomes. In contrast to viral sequences, however, mitogenomes have attracted considerable interest in the bioinformatics community, resulting in the development of a wide array of specialized tools (reviewed (Bernt *et al.*, 2012)). This software often capitalizes on the fact that the small size of the mitogenomes makes it possible to employ much more expensive algorithms than could be feasibly used in the context of prokaryotic or even eukaryotic genomics. Given the many specific questions and importance of RNA viruses in both basic research and in medical sciences it is hard to understand why the bioinformatics community has shown little interest in developing a comprehensive suite of methods and tools for RNA virology. The open problems remain many and diverse, ranging from orthology detection, protein annotation, and deep phylogeny to the evaluation of multiple, superimposed selection pressures, the evolution of viral gene regulation, and the understanding of the rapidly evolving populations of viruses and their arms race with the host immune system.

Acknowledgements

This work was supported in part by grants from the *Deutsche Forschungsgemeinschaft* (STA 850/7-2 to PFS, MI 439/14-1 to MMi and PFS, MA5082/1-1 to MMa), and by the Swiss National Science Foundation Grant no. CR32I2_127017. Furthermore, we thank Allan Clark for critically reading the manuscript.