From the Institute of Biochemistry
of the University of Lübeck
Director: Prof. Dr. Rolf Hilgenfeld

Analysis of overlapping reading frames in viral genomes

Dissertation in Fulfillment of the
Requirements for the Doctoral Degree
of the University of Lübeck

from the Department of Natural Sciences

Submitted by

Aditi Shukla

from Ranchi, India

Lübeck, 2015

1. First Referee: Prof. Dr. Rolf Hilgenfeld

2. Second Referee: Prof. Dr. Bernhard Haubold

Date of Oral Examination: 20.10.2015

Approved for printing: Lübeck, on 20.10.2015

Acknowledgements

I thank,

Prof. Hilgenfeld with deep gratitude for providing me with this research opportunity and making me a part of your esteemed Institute. I thank you for sharing your knowledge and ideas, providing support and encouragement throughout, and for an extensive review of this work. Your attention to detail towards every project, every proposal, every manuscript, amazes me. I wish to imbibe this immaculateness in my professional life.

Prof. Martinetz and Prof. Mamlouk for interesting discussions, useful critiques for this research work, and for helping to keep my progress on schedule.

The friendly management of the graduate school for your sustained support throughout. Special thanks to Olga and Katja for organizing interesting workshops, seminars and get-togethers.

Achim for your technical assistance; Mrs. Rosenfeld for making my initial arrival in Germany very easy and pleasant and Mrs. Schwab for work behind the scenes!

Raffaele, Monarin, Raspudin, Zhenggang, Linlin, Friedrich and Caroline for your kind fraternity; Jeroen and Ksenia for your suggestions and encouragements whenever required; for being there as a support! Other staff members at the institute for making the institute so interesting.

Upasana and Susmita for your friendship.

My family for immense patience, undying love and constant support.

Contents

Abstract

Virus evolution is subject to several restrictions, which include the rather small size of the viral genome and the necessity to replicate in a host cell. The latter implies, constantly trying to evade the host immune system. Hence, viruses must evolve very fast in order to adapt to changing environmental conditions within the hosts. Analysis of virus evolution is complicated by the frequent occurrence of overlapping reading frames. This phenomenon is observed in a number of RNA viruses of different genome sizes - from as small as that of HIV, to large ones such as those of coronaviruses. To understand the evolution of viral proteins created by overlapping reading frames, a systematic analysis of overlapping genes was carried out for SARS-coronavirus and for murine norovirus. The study involved codon-usage analysis, predicting disorder content in the overlapping proteins, complemented by structural studies wherever possible, prediction of the RNA secondary structure elements at the translational initiation sites of the alternative reading frames, and mutational analysis of the overlapping genes. Based on these analyses, it could be concluded that (1) usage of overlapping reading frames is one of the mechanisms employed by viruses for acquiring new protein domains, (2) overlapping protein regions have a tendency to be inherently disordered, (3) there is no consensus RNA secondary structure present at the translation initiation site of the alternative reading frame, which could assist in the non-canonical translation - instead, the RNA secondary structure appears to be case-specific for each overprinting protein -, and (4) viruses with overlapping reading frames might use a mechanism called “independent evolution” to escape any deleterious effect of mutations occuring in the overlapping gene set.

Abstract

Im Rahmen der Evolution unterliegen Viren einer Reihe von einschränkenden Rahmenbedingungen. In erster Linie spielt dabei die geringe Größe ihres Genoms eine Rolle, wodurch die Anzahl der verfügbaren Gene beschränkt ist. Darüber hinaus besteht für Viren die Notwendigkeit, Strategien zu entwickeln, um nicht nur in einer Wirtszelle erfolgreich replizieren, sondern auch das Immunsystems umgehen zu können. Aus den genannten Gründen müssen Viren eine sehr schnelle Entwicklung durchlaufen, um sich den sich ändernden Bedingungen im Wirtsorganismus erfolgreich anzupassen. Die Erforschung der Evolution von Viren ist somit von großem Interesse, stellt jedoch durch das häufige Auftreten von überlappenden Leserastern eine große Herausforderung dar. Dieses Phänomen kann bei einer ganzen Reihe von RNA-Viren unterschiedlicher Genomgröße, angefangen vom kleinen Genom des HI-Virus bis hin zu den größeren der Coronaviren, beobachtet werden. Um die Evolution viraler Proteine besser verstehen zu können, wurde eine systematische Untersuchung der Gene vorgenommen, die durch überlappende Leseraster codiert werden. Der Fokus lag beim SARS-Cornavirus sowie beim murinen Norovirus. Die Untersuchung beinhaltete die Nutzung der möglichen Codons sowie Vorhersagen über den Grad der Fehlordnung der überlappenden Proteine. Diese Vorhersagen wurden durch Strukturuntersuchungen und Mutationsstudien der überlappenden Bereiche vervollständigt. Darüber hinaus wurde die RNA auf das Vorhandensein von Sekundärstrukturelementen an den Stellen der Translationsinitiation alternativer Leseraster untersucht. Anhand der gewonnenen Ergebnisse konnten folgende Schlüsse gezogen werden: (1) Überlappende Leseraster stellen einen Mechanismus zum Erwerb neuer Proteindomänen durch Viren dar. (2)

Proteine, die durch überlappende Leseraster codiert werden, besitzen eine höhere Tendenz, fehlgeordnete Bereiche aufzuweisen. (3) An Stellen der Translationsinitiation alternativer Leseraster existiert kein Konsensus-RNA-Sekundärstrukturelement zur Unterstützung nicht-kanonischer Translation. Vielmehr scheint die vorliegende RNA-Sekundärstruktur spezfisch für die jeweiligen überlappenden Proteine zu sein. (4) Zur Vermeidung von schädlichen Mutationen in überlappenden Leserastern haben diese Viren vermutlich einen altenativen Mechanismus entwickelt, der auch als “unabhängige Evolution” bezeichnet wird.

Declaration

Results presented in this thesis have been partially published in Liu et al. (2014) and Shukla and Hilgenfeld (2015). The part of the Liu et al (2014) article on accessory proteins of non-SARS coronaviruses was written by me.

References:

1. Liu DX, Fung TS, Chong KK, **Shukla A,** Hilgenfeld R (2014): Accessory proteins of SARS-CoV and other coronaviruses. Antiviral Res. 109: 97-109.

2. **Shukla A,** Hilgenfeld R (2015): Acquisition of new protein domains by coronaviruses: analysis of overlapping genes coding for proteins N and 9b in SARS coronavirus. Virus Genes 50: 29-38.

1. Introduction

1.1 Overlapping reading frames

A codon is composed of three nucleotides and therefore, three reading frames are, in principle, possible within the same gene. The correct reading frame is determined by the start codon, usually ATG or, in RNA genomes, AUG, coding for methionine. If the same stretch of genome codes for two or more proteins in different reading frames, the open reading frames (ORFs) are called overlapping open reading frames or just overlapping reading frames. Proteins encoded by overlapping genes are often called "overlapping proteins" or "transframe proteins" (Plant, 2012), because during their translation, there occurs either a +1 or -1 shift in the reading frame (Fig. 1); (Belshaw et al., 2007).

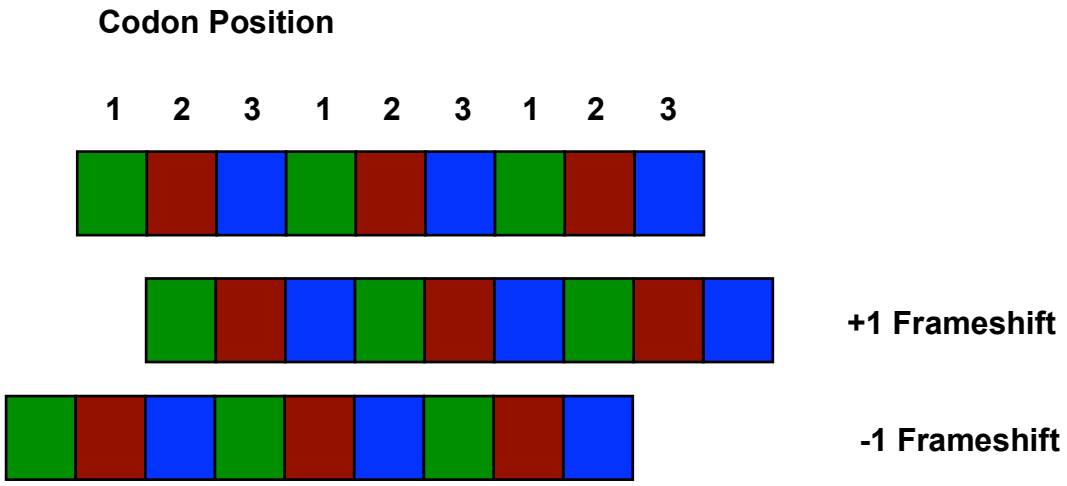


FIGURE 1 Types of frameshift with respect to the shift in codon position during translation of overlapping genes.

Overlapping genes were first described in the single-stranded DNA bacteriophage ϕX174 (Barell et al., 1976, Sanger et al., 1977). They are most

commonly found in viral genomes, where they are thought to be a result of evolutionary pressure so as to maximize its informational content while maintaining its small size (Miyata and Yasunaga, 1978; Krakauer, 2000). Overlapping genes are also found in eukaryotic genomes and are speculated to be involved in regulation of gene expression (Keese and Gibbs, 1992; Krakauer and Plotkin 2002).

Overlapping reading frames are translated into protein products that have been designated as “overprinted” (or “ancestral”) and “overprinting” (or “novel”), respectively (Rancurrel et al., 2009). Translation of overlapping reading frames in RNA viruses occurs via leaky ribosomal scanning (Kozak, 1989; 1999; 2002; Zou and Brown, 1996; Ryabova et al., 2006; Matsuda and Dreher, 2006), internal ribosomal entry site (IRES) (Jackson et al., 1990; Thiel and Siddell, 1994) programmed ribosomal frameshifting (Jacks et al., 1988; Brierley et al., 1989; Brierley and Dos Ramos, 2006; Dinman, 2012), or stop-codon readthrough (Beier, 1984; Honigman, 1991; Orlova, 2003).

1.1.1 Types of overlap

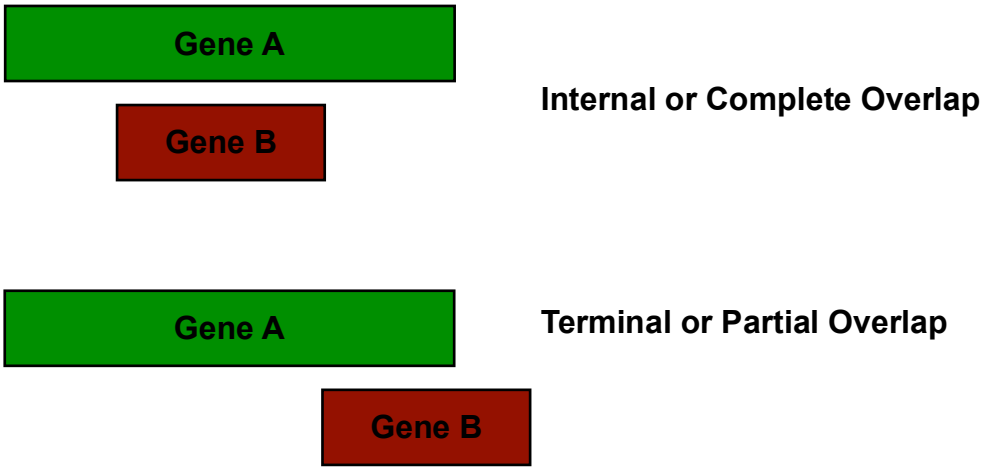


FIGURE 2 Types of overlap.

Depending on the extent of overlap, two types of the phenomenon can occur in overlapping genes. If the sequence of one gene overlaps completely with the sequence of another gene, the overlap is called “internal” or “complete”, and if only a part of the sequence of one gene overlaps with the sequence of another gene, then the overlap is called “terminal” or “partial” (Belshaw et al., 2007).

1.2 Direction and phase of overlapping reading frames

1.2.1 Direction of overlap

Genes can overlap either on the same strand, or, in case of a double-stranded genome, on the reverse complementary (antiparallel) strand. Hence, in principle, in an overlapping gene set, three directions of the overlap mode can occur: unidirectional (➔➔), convergent (➔➔), and divergent (➔➔) (Normark et al., 1983; Fukuda et al., 1999; 2003; Rogozin et al. 2002). In bacterial genomes, the majority (~84%) of the overlap occurs in the same direction of the reading frame (unidirectional) and not in the reverse direction (Johnson and Chrisholm, 2004). A systematic study on bacterial overlapping genes revealed that unidirectional overlap is the most common mode of overlap. Convergent overlaps are found to be not so common, whereas divergent overlaps are the rarest (Fukuda et al., 1999; 2003; Lillo and Krakauer, 2007). For example, there are 260 conserved overlapping genes in *Mycoplasma pneumoniae* and *Mycoplasma genitalium,* out of which 230 occur in the unidirectional mode, 28 occur in the convergent mode, and 2 occur in the divergent mode (Fukuda et al., 1999; 2003). There is also a

relationship between the direction of overlap and intergenic distances between the two overlapping genes. It has been shown that the average intergenic distance between the overlapping genes is least in the most common form of unidirectional overlaps, being 97 bp, whereas this distance is 144 bp in convergent, and 236 bp in divergent overlaps. It is presumed that the intergenic distance affects the frequency of the overlapping genes (Fukuda et al., 2003).

1.2.2 Phase of overlap in overlapping reading frames

In an overlapping gene set, the position of the codon in two reading frames with respect to each other is called the phase of the overlap (Krakauer 2000; Rogozin et al. 2002; Lillo and Krakauer, 2007; Johnson and Chrisholm 2008).

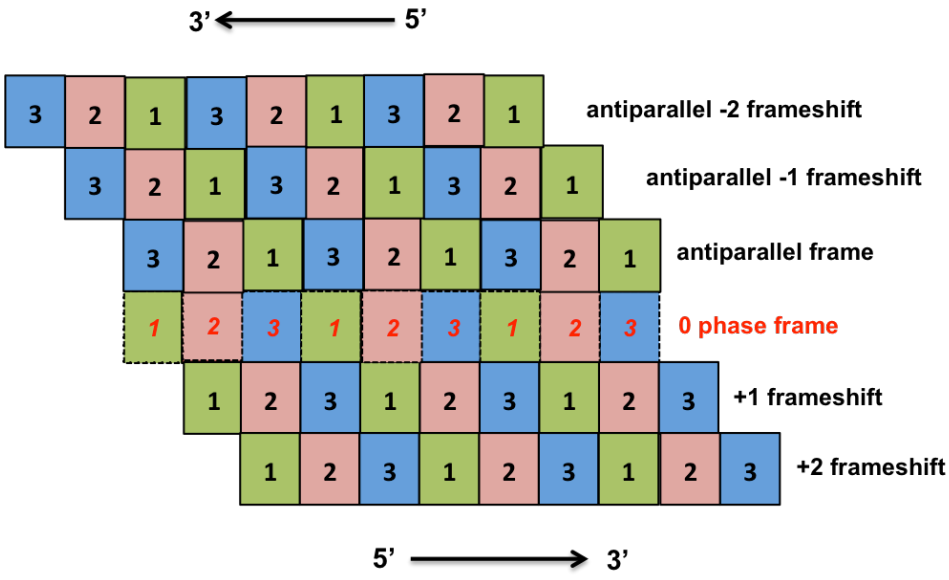


FIGURE 3 Direction and phase of overlapping genes (adapted and modified from Krakauer, 2000).

A unidirectional overlapping gene set can feature either a +1 frameshift or a +2 frameshift (Fig. 3). The +2 frameshift is also termed -1 frameshift (Belshaw

et al., 2007), because it produces the same effect on codon positioning in the dual-coding region. In terms of evolution, overlapping genes are considered a mechanism for creating novel proteins (Krakauer, 2000; Rancurrel et al., 2009). Any point mutation occurring in an overlapping gene region affects two (or more) protein products at the same time. Hence, the co-occurrence of a codon position in different reading frames, in other words the phase of the overlap, determines the effect of mutations in each reading frame.

1.3 Selection pressure on overlapping reading frames

1.3.1 Degeneracy of the genetic code

The standard genetic code (Fig. 4) is almost universal (Koonin and Novozhilov, 2009).

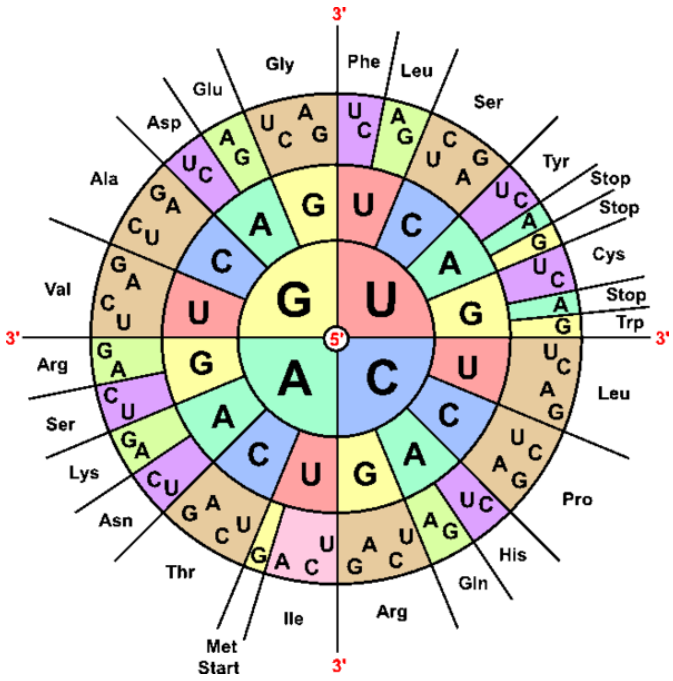


FIGURE 4 Standard genetic code (figure obtained from wikipedia page: http://de.wikipedia.org/wiki/Code-Sonne)

Since the last universal common anscestor (LUCA), the code has not changed much, apart from a few reassignments (Crick, 1968). However, in present days, there is enough evidence that the standard code is not literally universal, but there are 25 additional alternative genetic codes available in the NCBI taxonomy database depicting genetic codes (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=tgencodes#SG1, last updated on April 30th 2013, last accessed on 20th February 2015). Nevertheless, it has to be emphasized that the modifications do not alter the basic standard organization (Koonin and Novozhilov, 2009).

Most of the 20 amino acids (except for methionine and tryptophan) are coded by more than one codon. The codons which code for the same amino acid are called synonymous codons. Upon a point mutation in a coding region of a genome, the position at which nucleotide substitution occurs within a codon reflects whether the substitution would be synonymous or not. When there is a nucleotide substitution at the first codon position, it causes an amino-acid change in 60 out of 64 cases. The four exceptions, two each in codons for leucine (UUA-CUA, UUG-CUG) and arginine (AGA-CGA, AGG-CGG), occur because of the partial degeneracy of codons (Fig. 4). When there is a nucleotide substitution at the second codon position, it results in an amino-acid change in 63 out of 64 cases. At this codon site, the only substitution that is synonymous occurs in stop codons (UAA-UGA). Lastly, a nucleotide substitution occurring at the third codon position causes a change in amino acid in only 16 out of 64 cases because of codon degeneracy at the third nucleotide position (Fig. 4) (Pavesi et al., 1997; 2006; Matsuda and Dreher, 2006).

1.3.2 Synonymous and non-synonymous substitutions

Synonymous substitutions result in silent mutations and non-synonymous substitutions result in non-silent mutations. The ratio ω of non-synonymous (K_a) to synonymous (K_s) nucleotide substitution rates is an indicator of selective pressure on genes. A ratio significantly greater than 1 indicates positive selective pressure. A ratio around 1 indicates either neutral evolution at the protein level or an averaging of sites under positive and negative selective pressure. A ratio less than 1 indicates pressure to conserve protein sequence, i.e. “purifying selection” (Hurst, 2002).

1.3.3 Different kinds of evolutionary strategy adopted by overlapping genes

a) Evolution by directional (positive) selection in one reading frame while purifying (negative) selection in the other: This is the most common mechanism seen in overlapping gene sets (Miyata and Yasunaga, 1978; Pavesi, 2000; 2006; Jordan et al., 2000; Fujii et al. 2001).

b) Evolution where both the reading frames are undergoing directional selection (Rogozin et al., 2002):

c) Evolution where both reading frames evolve neutrally or favoring just synonymous substitutions so that the physico-chemical properties of the translated protein products are not changed: This mechanism is also called “Constrained Evolution” (Mizokami et al., 1997).

d) Evolution where one reading frame evolves neutrally or almost neutrally, while there is a positive or negative selection in the other reading frame: this mechanism is also called “Independent Evolution” (Zaaijer et al., 2007).

1.4 Viruses with overlapping reading frames

Although overlapping reading frames have been discovered in many organisms, they are most commonly found in viruses because of their comparatively small-sized genomes (Keese and Gibbs, 1992; Belshaw et al., 2007; Chirico et al., 2010). An evolutionary study on multiple overlapping genes within the genomes of the families *Rhabdoviridae* and *Paramyxoviridae* revealed negative selection in one protein product and concomitant rapid evolution of the other (Jordan et al., 2000). This kind of differential selection in overlapping genes, i.e. adaptive evolution in one reading frame and negative selection in the other, is quite common (Miyata and Yasunaga, 1978; Pavesi, 2000; 2006; Jordan et al., 2000; Fujii et al. 2001). Hepatitis B Virus (HBV) has two-thirds of its genome coding for more than one protein. A comparison of the evolutionary rates in overlapping and non-overlapping gene regions of HBV provided evidence of “constrained evolution”, i.e. the non-overlapping gene regions evolve faster than the overlapping gene regions (Mizokami et al., 1997). However, more recently, another study on the overlapping polymerase (P) and surface protein (S) genes of HBV was performed and “independent evolution” of both overlapping genes has been reported (Zaaijer et al., 2007).

1.5 Objective of this study

Many of the overlapping proteins in viruses are accessory or group-specific proteins. Some of them are known to contribute greatly to enhancing the pathogenicity of the virus. In the genome of Severe Acute Respiratory

Syndrome Coronavirus (SARS-CoV), several of the accessory genes overlap with other open reading frames (Rota et al., 2003; Marra et al., 2003; Narayanan et al., 2008; McBride and Fielding, 2012; Liu et al., 2014). The most interesting example among these is the accessory protein encoded by open reading frame (orf) 9b, which is an internal overlap with the N-terminal domain of the nucleocapsid protein (Marra et al., 2003; Rota et al., 2003). This is one of the rare cases in which crystal structures are available for both of the overlapping proteins. Also, a large body of sequence information is available for SARS-CoV isolates and SARS-like coronaviruses from bats and civets, allowing a study into the evolution of the overlapping N and 9b genes and characterization of the sequence properties of this gene set (Shukla and Hilgenfeld, 2015). Similarly, proteins encoded by orf3a and orf3b of SARS-CoV are another set of overlapping proteins which has a partial overlap of sufficient length to make meaningful sequence analysis (Marra et al., 2003; Rota et al., 2003). Also included in this study is the recently discovered overlapping accessory gene in Murine Norovirus, i.e. the Virulence factor 1 gene (Vf1) (McFadden et al., 2011). For comparison, the overlapping set of the genes coding for Hepatitis B virus polymerase and surface protein is included.

In this theoretical analysis, I have tried to answer the following questions:

(1) Is one of the gene products more ancient than the other, suggesting that overlapping orfs may be a mechanism for the virus to acquire new proteins?

(2) Do the overlapping sets of proteins tend to be inherently disordered? Are the three-dimensional structures (if available) of the overlapping proteins

optimized in terms of packing and stability, or is one of them (or both) making compromises?

(3) Is there a consensus RNA secondary structure element that is responsible for creation of overprinting proteins in viral genomes?

(4) How do mutations in one reading frame affect the protein encoded by the other reading frame, and vice versa?

(5) Do overlapping reading frames impose restrictions on mutations in the overlapping protein set, or are they evolving independently of each other?

2. Biological background

Accessory genes are group-specific genes usually seen to be associated with viral pathogenecity (Liu et al., 2014). To characterize sequence properties of newly acquired genes formed by overlapping reading frames, overlapping accessory genes and their respective protein products in Severe Acute Respiratory Syndrome (SARS) coronavirus and Murine norovirus (MNV) were studied. The obtained results were also compared with the most studied case in this field, Hepatitis B virus (HBV).

2.1 Accessory proteins of coronaviruses

Coronaviruses are positive-stranded RNA viruses which belong to the subfamily *Coronavirinae* in the family *Coronaviridae,* within the order *Nidovirales* (Fauquet et al., 2005). Coronaviruses are divided into three genera: ***Alpha-, Beta-,*** and ***Gamma*coronaviruses.** The *Betacoronaviruses* are further classified into lineages a, b, c, and d. A fourth genus, ***Deltacoronavirus,*** has been proposed in order to include new coronavirus species (de Groot and Gorbalenya, 2010). This fourth genus comprises a number of recently identified avian and a few mammalian coronaviruses (Woo et al., 2010; 2014).

According to the species demarcation criterion for the viral family *Coronaviridae*, viruses that share more than 90% amino-acid sequence identity in the conserved replicase domains are considered to belong to the same species (9th report of the International Committee on the Taxonomy of

Viruses , de Groot et al., 2011). If this new classification system is going to be accepted, bat coronaviruses will dominate the *Alpha-* and *Betacoronavirus* genera and avian coronaviruses the *Gamma-* and *Deltacoronavirus* genera (Woo et al., 2010; 2014; de Groot et al., 2011).

Many small open reading frames (ORFs) occurring in inconsistent numbers are present in coronaviruses downstream to ORF 1. These ORFs encode proteins which have not been previously characterized and hence, for most of them the functions are unknown. Deletion studies are commonly carried out with these proteins to identify if they are dispensable or indispensable for the viral life cycle. Coronaviruses are responsible for causing common cold in humans and are usually not life-threatening, with the exception of SARS coronavirus (SARS-CoV) and the newly identified Middle East Respiratory Syndrome (MERS) coronavirus (MERS-CoV). Apart from humans, coronaviruses infect many mammals such as bats, cattle, cats, pigs, horses, whales, and birds such as munia, thrush, bulbul etc.

A phylogenetic tree is shown below, for the members of all four coronavirus genera (Fig.5). In Appendix I, they are also listed along with the respective hosts, genome length as well as number and length of accessory proteins which they encode.

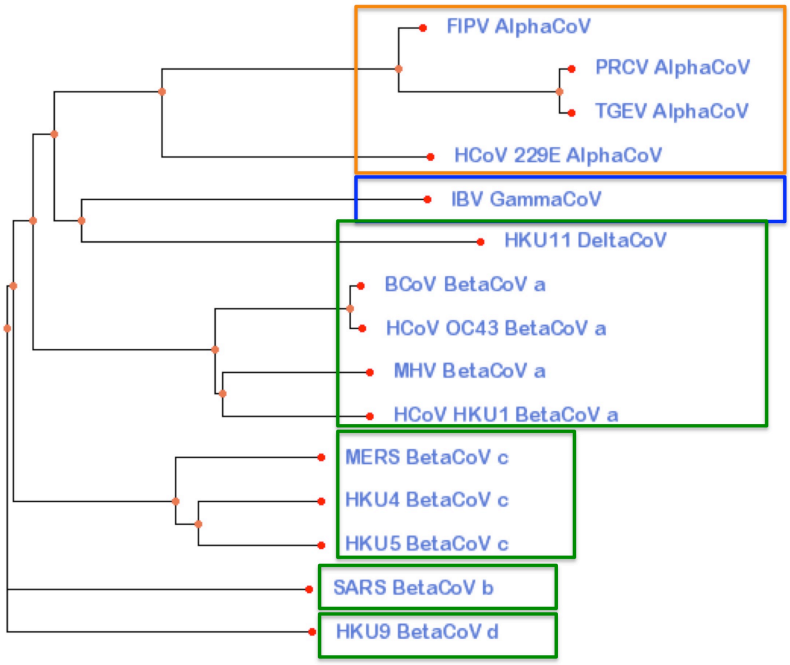


FIGURE 5 Phylogenetic analysis of the four genera of coronaviruses based on pp1ab using Kalign multiple sequence alignment and Phyfi tree viewer (Lassmann and Sonnhammer, 2005; Fredslund, 2006). Next to the branch, virus names along with the lineages are indicated. The tree highlights four main clusters corresponding to the genera *Alpha-*, *Beta-*, and *Gamma-CoVs* and the recently proposed new genus *Delta-CoV*. The four distinct betacoronavirus lineages a, b, c, and d can also be seen.

2.1.1 Alphacoronaviruses

Some of the most studied accessory genes among the members of *alphacoronavirus* are described in what follows:

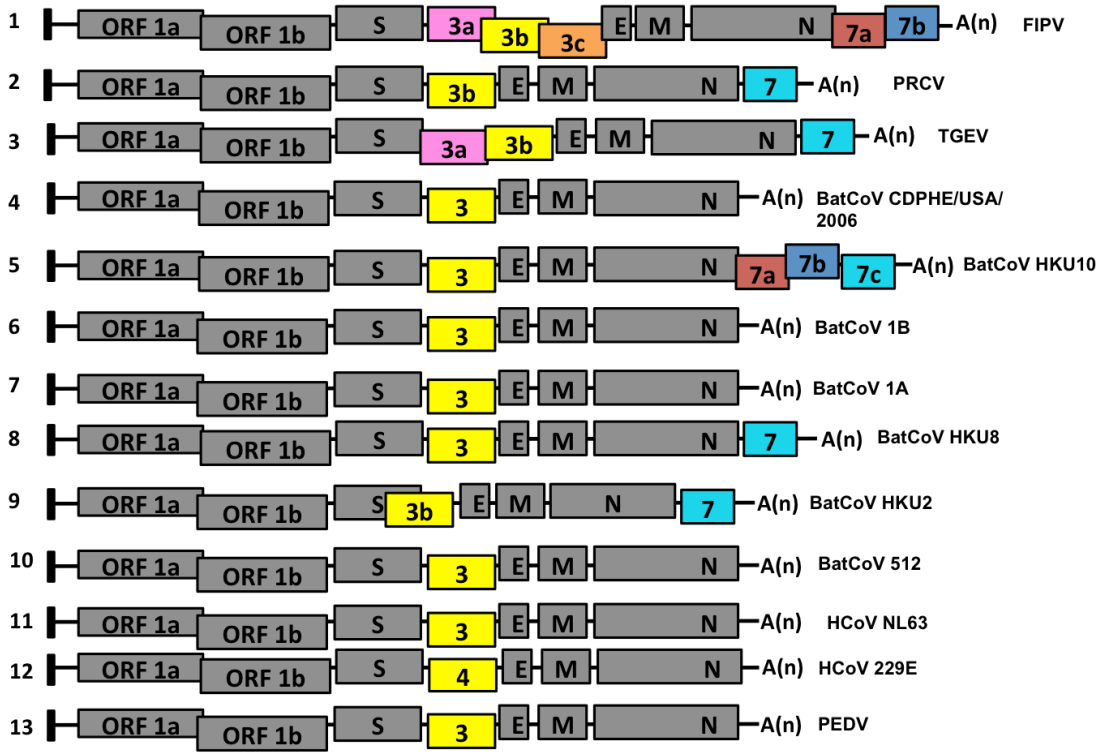


FIGURE 6 Genome organization of members of the genus *Alphacoronavirus* highlighting the accessory genes downstream of the orf1a/1b gene. Spike (S), envelope (E), membrane (M), and nucleocapsid (N) are the structural genes and interspersed between them are the genes coding for putative accessory proteins. 1. Feline infectious perotonitis virus (NC_002306), 2. Porcine respiratory coronavirus (DQ811787), 3. Transmissible gastroenteritis virus (DQ811788), 4. Bat coronavirus CDPHE 15/USA/2006 (NC_022103), 5. *Rousettus* bat coronavirus HKU10 (NC_018871), 6. *Miniopterus* bat coronavirus 1B (NC_010436), 7. *Miniopterus* bat coronavirus 1A (NC_010437), 8. *Miniopterus* bat coronavirus HKU8 (NC_010438), 9. *Rhinolophus* bat coronavirus HKU2 (NC_009988), 10. *Scotophilus* bat coronavirus 512 (NC_009657) 11. Human coronavirus NL63 (NC_005831), 12. Human coronavirus 229E (NC_002645), 13. Porcine epidemic diarrhea virus (NC_003436).

Feline infectious peritonitis virus (FIPV) is known to possess two genes downstream of the N gene (Fig. 6) . It has been demonstrated that these

ORFs are essential for efficient replication *in vitro* and for virulence *in vivo* (Haijema et al. 2004, Dedeurwaerder et al., 2013). Recent studies on the orf3a,b,c and orf7a,b accessory proteins of FIPV showed that the orf7 proteins are crucial for FIPV replication in monocytes/macrophages, giving an explanation for their importance *in vivo*. The orf3 proteins were found to have only supportive roles during the FIPV infection of the target cell (Dedeurwaerder et al., 2013). Investigations on the role of orf7 proteins in the evasion of the interferon (IFN)-mediated immune response indicated that the FIPV orf7a protein is a type-I IFN antagonist and protects the virus from the antiviral state induced by IFN, however, that it needs the presence of orf3 proteins to exert its antagonistic function (Dedeurwaerder et al., 2014). The orf3 proteins of Porcine Respiratory coronaviruses (PRCV) and Transmissible Gastroenteritis Virus (TGEV) are known to play potential roles in determining virulence (Tung et al., 1992; Paul et al., 1997). Another notable accessory protein of an alphacoronavirus is the protein encoded by orf4 in human coronavirus 229E. It has been previously known that the genome of human coronavirus 229E encodes two accessory proteins, namely orf4a and orf4b. However, recent complete genome sequencing of clinical isolates shows the presence of a full-length orf4 protein, whereas laboratory strains show the presence of a truncated orf4 (Farsani et al., 2012). It is suggested that extensive culturing of HCoV 229E might have resulted in this truncation. The highly conserved amino-acid sequence of the orf4 protein among clinical isolates suggests that the protein plays an important role *in vivo* (Dijkman et al., 2006).

2.1.2 Betacoronaviruses

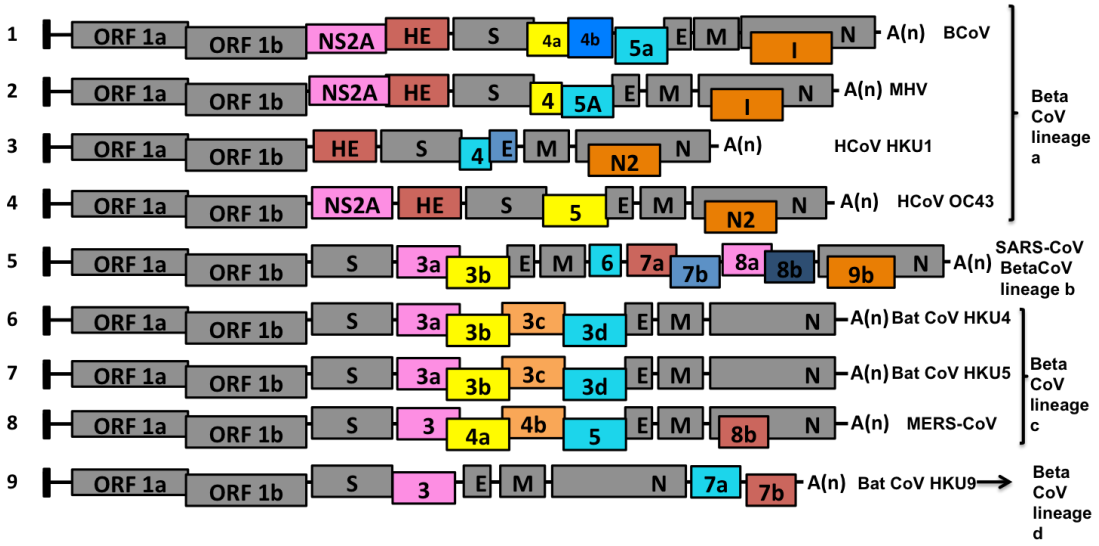


FIGURE 7 Genome organization of members of *Betacoronavirus* lineages a, b, c, and d, highlighting the accessory genes downstream of the orf1a/1b gene. Spike (S), envelope (E), membrane (M), and nucleocapsid (N) are the structural genes and interspersed between them are the genes coding for putative accessory proteins. 1. Bovine Coronavirus (NC_003045), 2. Mouse Hepatitis Virus (AC_000192), 3. Human Coronavirus HKU1 (NC_006577), 4. Human Coronavirus OC43 (NC_005147), 5. Severe Acute Respiratory Syndrome virus (NC_004718), 6. *Tylonycteris* bat coronavirus HKU4 (NC_009019),7. *Pipistrellus* bat coronavirus HKU5 (NC_009020), 8. MERS-CoV/Human Coronavirus 2c EMC/2012 (NC_019843), 9. *Rousettus* bat coronavirus HKU9 (NC_009021).

In ***betacoronaviruses lineage a***, for example in Bovine Coronavirus (BCoV), Mouse Hepatitis Virus (MHV), Human Coronavirus HKU1, Human Coronavirus OC43 etc., there occurs a haemaglutinin esterase (HE) gene between orf1ab and the S gene encoding a glycoprotein with neuraminate O-acetyl esterase activity (Fig. 7). This gene is exclusively present in betacoronaviruses of lineage a, suggesting its acquisition after diverging from

the anscestors of other betacoronavirus lineages. It has presumably been acquired via horizontal gene transfer from influenza C virus (Luytjes et al., 1988; Zhang et al., 1992; Zeng et al., 2008; Langereis et al., 2012). In addition to the HE accessory gene, studies have been performed on the internal ORF within the nucleocapsid gene of MHV. This ORF encodes a structural protein of 136 amino-acid residues that, however, is found to be non-essential for viral replication in either cell culture or in its natural host (Fischer et al., 1997).

Betacoronaviruses lineage b**,** for example SARS-CoV, is discussed later in this chapter (section 2.1.5).

In ***betacoronaviruses lineage c***, for example in the bat coronaviruses HKU4, HKU5, and the recently characterized MERS-CoV, there are four ORFs between the S and the E gene which encode the putative orf3a (also known as orf3), orf3b (also known as orf4a), orf3c (also known as orf4b) and orf3d (also known as orf5) proteins. MERS-CoV also contains an internal ORF within the nucleocapsid gene, which has not yet been characterized (van Bohemeen et al., 2012). This ORF was not previously described for BtCoV-HKU4 and BtCoV-HKU5 but is conserved in the genome sequences of both viruses. Transmembrane prediction revealed that the orf3 and orf4b proteins each contain one transmembrane helix, whereas the orf5 protein contains three such helices. It was recently suggested that the orf3 and orf5 proteins localize to the endoplasmic reticulum-golgi intermediate compartment (Niemeyer et al., 2013). In another study, the orf4b protein was found to be

localized to the nucleus (Matthews et al., 2014). Several studies have recently demonstrated that the orf4a, orf4b and orf5 proteins act as interferon antagonists with the orf4a protein being the most potent to counteract the antiviral effects of IFN via the inhibition of both interferon production and ISRE promoter element signaling pathways (Yang et al., 2013; Niemeyer et al. 2013). Moreover, the orf4a protein has recently been described as a double-stranded (ds) RNA-binding protein, which supresses the innate immune response by targetting the cellular dsRNA-binding protein PACT. It prevents the dsRNA intermediate products of viral RNA replication from binding to PACT, resulting in failure of activation of the cellular dsRNA sensors RIG-I and MDA5 (Siu et al., 2014).

In ***Betacoronavirus lineage d,*** for example in bat coronavirus HKU9, two accessory proteins, orf7a and orf7b, are found downstream of the N gene (Fig. 7) (Woo et al., 2007; Lau et al., 2010). These are encoded by two ORFs, both of which are connected with a transcription regulatory sequence (TRS). A TRS is a highly conserved AU-rich core sequence located at the 5′ end of coronavirus genes, which is essential for mediating a 100- to 1,000-fold increase in mRNA synthesis when located in the appropriate context (Alonso et al., 2002; Sawicki et al., 2007). Thus, the identification of TRSs preceding a predicted ORF helps identify new genes. This is the first time in a betacoronavirus that ORFs downstream of the N gene have been observed.

2.1.3 Gammacoronaviruses

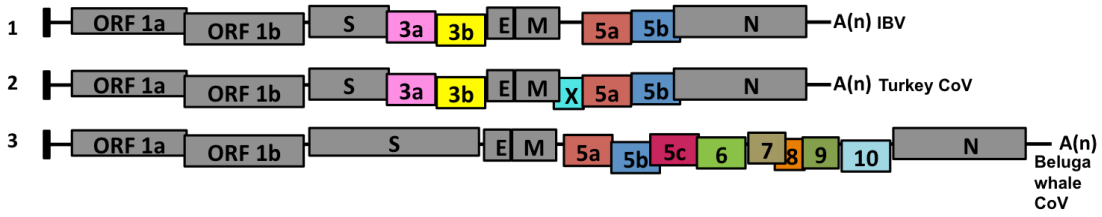


FIGURE 8 Genome organization of members of the genus *Gammacoronavirus* highlighting the accessory genes downstream of the orf1a/1b gene. Spike (S), envelope (E), membrane (M), and nucleocapsid (N) are the structural genes, and interspersed between them are the genes coding for putative accessory proteins. 1. Infectious Bronchitis Virus (NC_001451), 2. Turkey coronavirus (NC_010800), 3. Beluga whale coronavirus SW1 (NC_010646).

The most studied gammacoronavirus to date is Infectious Bronchitis Virus (IBV). The genomic organization of the classic gammacoronaviruses is as follows:

5′-Pol-S-3a-3b-E-M-5a-5b-N-(UTR)-3′

It contains four accessory genes coding for four accessory proteins, namely orf3a, orf3b, orf5a, and orf5b (Fig. 8). Reverse genetics studies have shown that these are dispensable for viral replication (Casais et al., 2005; Hodgson et al., 2006).

2.1.4 Deltacoronaviruses

Recent complete genome sequencing and comparative analysis showed that both the avian and mammalian coronaviruses belong to the genus *Deltacoronavirus*, with similar genome characteristics and structures (Woo et al., 2009; 2012).

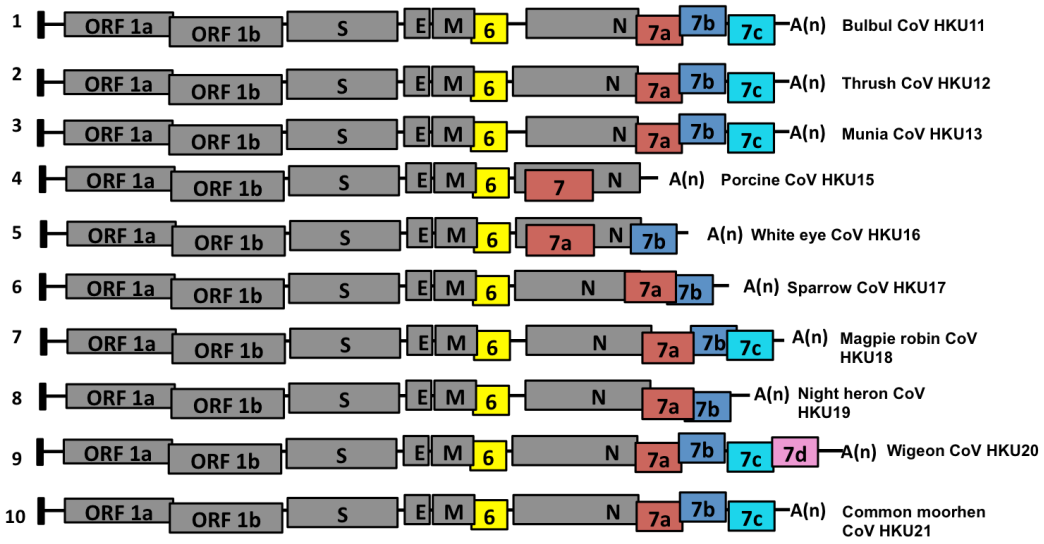


FIGURE 9 Genome organization of members of the genus *Deltacoronavirus* highlighting the accessory genes downstream of the orf1a/1b gene. Spike (S), envelope (E), membrane (M), and nucleocapsid (N) are the structural genes and interspersed between them are the genes coding for putative accessory proteins. 1. Bulbul coronavirus HKU11 (FJ376620), 2. Thrush coronavirus HKU12 (NC_011549), 3. Munia coronavirus HKU13 (NC_011550), 4. Porcine coronavirus HKU15 (JQ065042), 5. White eye coronavirus HKU 16 (JQ065044), 6. Sparrow coronavirus HKU17 (JQ065045), 7. Magpie robin coronavirus HKU18 (JQ065046), 8. Night heron coronavirus HKU19 (JQ065047), 9. Wigeon coronavirus HKU20 (JQ065048), 10. Common moorhen coronavirus HKU21 (JQ065049).

Their genome size is the smallest amongst all coronaviruses, ranging from 25.4 to 26.6 kb. Currently, all the 10 complete genome sequences possess an orf6 accessory gene located between the M and N genes, and a variable number (one to four) of accessory genes downstream of the N gene (Woo et al., 2009; 2012) (Fig. 9). These group-specific genes are yet to be characterized.

2.1.5 SARS-coronavirus: A member of betacoronavirus lineage b; genome organization and its overlapping proteins

SARS is an acute respiratory illness characterized by pulmonary inflammation with a fatality rate of about 10% (Hilgenfeld and Peiris, 2013; Cheng et al., 2013). SARS-CoV first emerged in 2002 in Guangdong province, China, and by the spring of 2003 developed into a widespread epidemic (Cheng et al., 2013). The origin of this virus is believed to be a bat reservoir (Li et al., 2005; Drexler et al., 2014; Ge et al.; 2013). Although the epidemic ended in July 2003, it is not unlikely for SARS-CoV or a similar virus to re-surface again. This view is supported by the recent emergence of MERS-CoV, a new human betacoronavirus of lineage c (Zaki et al., 2012; de Groot et al. 2013; van Bohemeen et al. 2012; Cotten et al. 2013). The primary reservoir of MERS-CoV has been shown to be dromedary camels (Reusken et al., 2013; Meyer et al., 2014, Haagmans et al., 2014), however, the possibility of an origin from bats cannot be ruled out (Annan et al., 2013; Ithete et al., 2013; Yang et al., 2014). As of April 9th, 2015, 1102 laboratory-confirmed human MERS cases have been reported since September 2012, including 416 deaths (http://www.who.int/csr/don/9-april-2015-mers-saudi-arabia/en/).

The emergence of yet another life-threatening coronavirus emphasizes the importance of understanding the pathogenesis of these viruses.

The genome of SARS-CoV codes for at least 28 proteins. Two-thirds of the 30-kb single-stranded RNA genome of positive polarity comprise ORF 1ab, which encodes the viral polyproteins pp1a and pp1ab. The 3'-proximal third comprises orfs encoding the structural proteins, i.e. spike (S), envelope (E),

membrane (M), and nucleocapsid (N) (Marra et al., 2003; Rota et al., 2003). In addition, several small ORFs coding for the accessory proteins are distributed among the gene segments coding for the structural proteins; these include orf3a/b, orf6, orf7a/b, orf8a/b, orf9b and possibly orf9c (Narayanan et al. 2008; Mc Bride and Fielding, 2012; Liu et al., 2014). SARS-CoV accessory proteins are thought to play important roles in viral pathogenicity (Tan et al., 2006; Narayanan et al. 2008; Mc Bride and Fielding, 2012; Liu et al., 2014). The present work focuses on the internal overlap of the orf9b gene within the nucleocaspsid gene and on the partial overlap in the orf3a and orf3b accessory genes (Fig. 10).

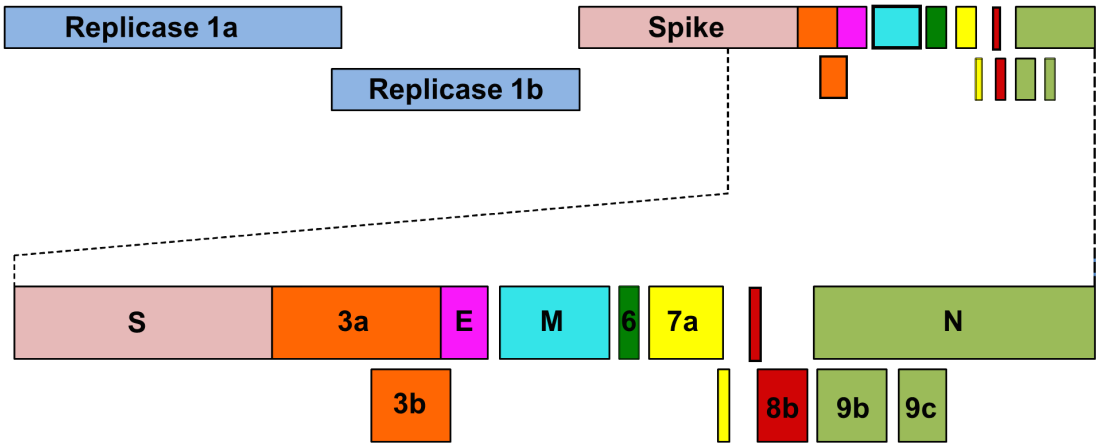


FIGURE 10 Schematic organization of the SARS-coronavirus genome, highlighting structural and accessory genes. The overprinting accessory genes are indicated below their overprinted mates.

Internal overlap of the orf9b gene :

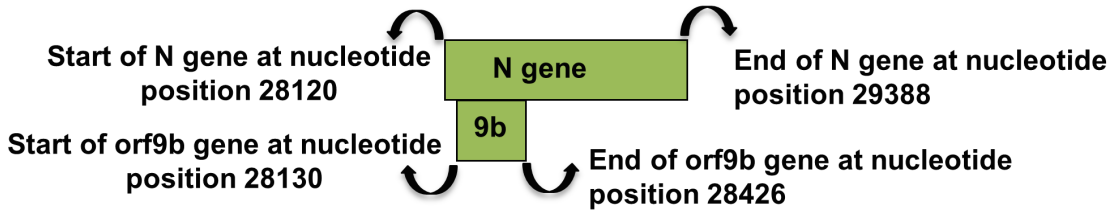


FIGURE 11 Overlapping N and orf9b genes of SARS-CoV with their start-stop co-ordinates in its genome ("end" coordinates include the stop codon).

The orf9b gene completely overlaps with the nucleocapsid gene. During the translation of the subgenomic mRNA coding for the nucleocapsid protein, a +1 frameshift may occur at the 10th nucleotide, leading to the translation of orf9b (Fig. 11). Recently, it has been proposed that leaky ribosomal scanning is responsible for the production of the orf9b protein (Xu et al., 2009). Orf9b codes for a small protein of 98 amino-acid residues, which is present in SARS-CoV-infected cells (Chan et al., 2005). Antibodies against this protein have been detected in the sera of SARS patients, demonstrating that the protein is produced during infection (Qiu et al., 2005). Its three-dimensional structure has been determined by X-ray crystallography, revealing a previously unknown fold (Meier et al., 2006). The orf9b protein is thought to be a virion-associated protein (Xu et al., 2009). It has also been postulated that it possesses a nuclear export signal (NES) and is localized to the extra-nuclear region (Moshynskyy et al., 2007). Recently, it has been found that the orf9b protein diffuses into the nucleus, undergoes active Crm1-mediated nucleocytoplasmic export and, when retained in the nucleus, triggers apoptosis (Sharma et al., 2011). However, the exact function of the orf9b

protein during SARS-CoV infection is not well understood. Also, there is no sequence similarity between the orf9b protein and any other known protein (Marra et al., 2003, Rota et al., 2003). Among other lineages of *Betacoronaviruses,* internal ORFs within the nucleocapsid gene have been reported, but these so-called "internal genes" have little in common with the orf9b gene of SARS-CoV (Liu et al., 2014). For the purpose of comparison, they were also included in this work (Fig. 12).

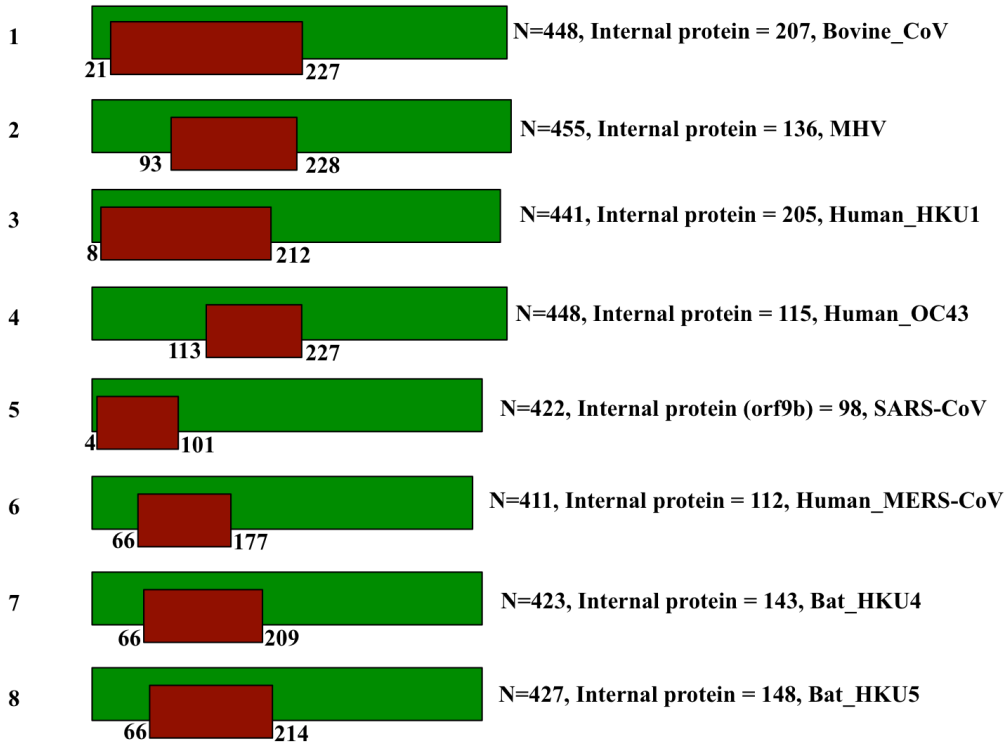


FIGURE 12 Internal overlapping proteins within the nucleocapsid protein in members of the genus Betacoronavirus. Protein lengths (number of amino-acid residues) are indicated next to each overlapping protein set.

In case of SARS-CoV, there has also been a report of another overlapping gene, orf9c (nucleotides 28,583 to 28,795) within the nucleocapsid gene (see Fig. 10). It encodes a predicted protein of 70 amino-acid residues. BLAST

analysis failed to identify similar sequences. A single trans-membrane helix was predicted using the TMpred (Trans Membrane prediction) algorithm (Marra et al., 2003). Until now, no evidence of orf9c expression has been found and since it is not yet annotated, it is not included in this study.

Partial overlap of the orf3a and orf3b genes:

The orf3a gene of SARS-CoV codes for a protein comprising 274 amino-acid residues and the orf3b gene codes for a protein of 154 amino-acid residues. There is a partial overlap of 134 amino-acid residues (Fig. 13). Both the orf 3a and orf3b protein are not homologous to any other known proteins (Marra et al., 2003).

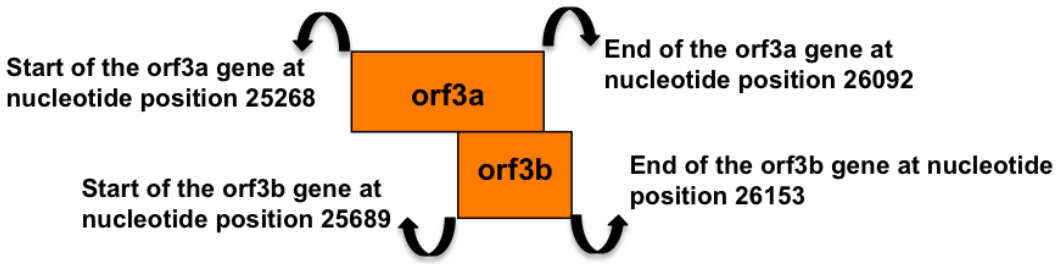


FIGURE 13 Co-ordinates of the orf3a and orf3b genes in SARS-CoV ("end" coordinates include the stop codon).

The orf 3a protein was previously also known as X1 or the U274 protein (Rota et al., 2003; Tan et al., 2004). Expression of the orf3a gene was detected, both *in vitro* in SARS-CoV infected cells, and *in vivo* in the lungs of a SARS patient (Yu et al., 2004; Guo et al., 2004; Zeng et al., 2004). This protein was found to localize to the perinuclear region as well as to the plasma membrane (Tan et al., 2004; Ito et al., 2005). Ito et al. (2005) found that the orf3a

protein is virion-associated and hence, might be a structural protein. This was also confirmed by Shen et al. (2005), although it was shown that its function is dispensible. The orf3a protein is thought to modulate virus release through its potassium-sensitive channel function (Lu et al. 2006; Chan et al. 2009). Moreover, it is also known to regulate host cellular responses, e.g. by inducing apoptosis (Law et al., 2005; Zhong et al., 2006; Chan et al. 2009; Freundt et al., 2010). It has been shown that frameshift mutations in the orf3a gene are responsible for the formation of multiple orf3a variant proteins, each having a different length (Tan et al., 2005; Wang et al., 2006). In some patients, only the truncated orf3a protein forms were found, and so it is speculated that each variant may have a different stability or different function and might contribute differently to the viral pathogenesis *in vivo* (Tan et al., 2005).

The orf 3b protein was previously also known as X2 protein or U154 (Rota et al., 2003; Tan et al., 2004). Expression of the orf3b gene was detected, both *in vitro* and *in vivo* (Guo et al., 2004; Chan et al., 2005). It was shown that the orf3b protein localizes to the nucleolus of infected Vero E6 cells (Yuan et al., 2005). Later, it was also seen to localize to mitochondrial cells due to the presence of a nuclear export sequence (Yuan et al., 2006; Freundt et al., 2009). The orf3b protein is reported to induce apoptosis and necrosis in the infected cells (Khan et al., 2006) and is thought to function as an interferon antagonist through inhibition of IRF3 (Kopecky-Bromberg et al., 2007).

Till date, there are no crystal structures available for either the orf3a or orf3b proteins (Hilgenfeld and Peiris, 2013).

2.2 Overlapping protein in Murine Norovirus

Murine norovirus (MNV) is a species of *norovirus* affecting mice. It was first identified in 2003 in immunocompromised mice (Karst et al., 2003). Like other members of the *Caliciviridae* family, MNV has a positive-sense, single-stranded RNA genome.

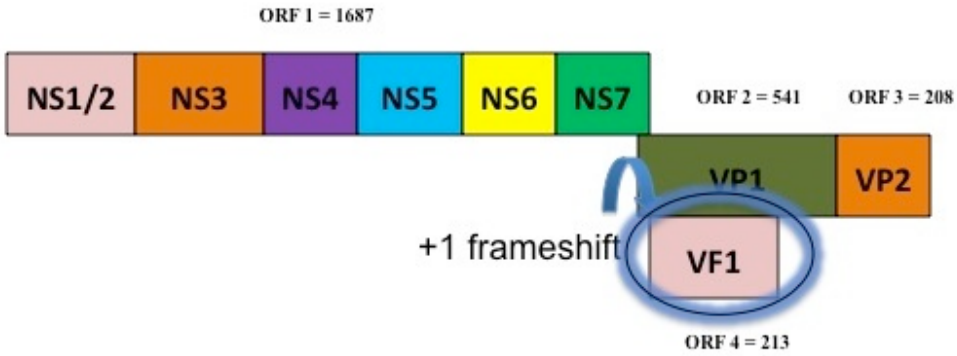


FIGURE 14 Schematic organization of the MNV genome.The overprinting accessory gene Vf1 is indicated below its overprinted mate, VP1.

However, unlike the other members of the *Caliciviridae,* which comprise three orfs, it has been recently detected that MNV encodes a potential alternative open reading frame coding for an accessory protein annotated as Virulence factor 1 (Vf1) protein (Fig. 14). This Vf1 protein is formed by a +1 shift at the 13th nucleotide position of the ORF coding for the VP1 protein. The VP1 protein comprises 541 amino-acid residues and there is a complete overlap with Vf1, which comprises 213 amino-acid residues.

2.3 Overlapping proteins in Hepatitis B Virus (HBV)

In studies of overlapping reading frames, the genome of HBV is extensively used. Twp thirds of the HBV polymerase (P) gene contain different alternating

reading frames. HBV has a circular genome comprising 3215 nucleotides and coding for 7 proteins (Fig. 15).

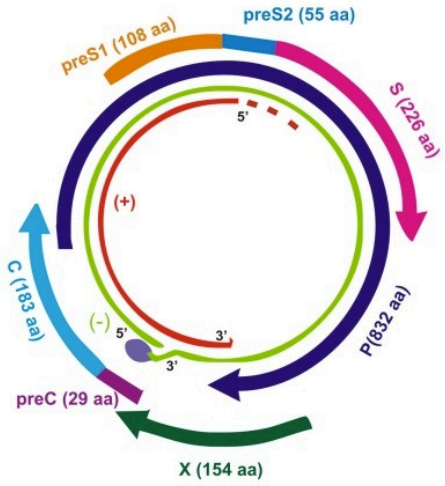


FIGURE 15 Schematic organization of the partially double-stranded HBV genome showing two-thirds of the genes overlapping each other (Figure adapted from Jayalakshmi et al. 2013).

Comprising 226 amino-acid residues, the surface (S) protein of HBV overlaps completely with the P protein, which comprises 843 amino-acid residues. The remaining 5 proteins partially overlap with each other (Fig. 15). A number of studies have been carried out on evolutionary aspects of these alternative reading frames (Mizokami et al., 1997; Zaaijer et al., 2007; Chen et al., 2013).

3. Special sequence properties of overlapping genes

Various studies have analyzed the information content in overlapping gene regions (Pavesi et al. 1997; 2013). To characterize the special sequence properties of genes involved in overlapping reading frames, comparisons of the patterns of codon usage in overlapping and non-overlapping genes are carried out.

3.1 Codon usage bias

Due to the degeneracy of the genetic code, 18 out of 20 different amino acids are encoded by more than one codon, the exceptions being methionine and tryptophan (for details, see Chapter 1). The frequencies of occurrence of alternative synonymous codons are not equal in a gene, therefore leading to a "codon usage bias" in that gene. The phenomenon has been studied since the beginning of nucleic acid sequence determination and sequence database creation (Grantham et al., 1980; 1981). These studies resulted in the generation of genome hypotheses which state that genomes within a species or related species have a similar pattern of codon usage. In other words, each genome within a species shows uniformity in using some synonymous codons preferentially over others (Grantham et al., 1980; 1981; Ikemura, 1985; Sharp and Li, 1987, Sharp et al., 2010). It was also shown that adding up the codon usage pattern of all the genes in an organism to get the total codon usage of the organism may conceal the underlying heterogeneity and hence, it is better to specify the codon usage trends among the genes in a species and between closely related species (Aota and Ikemura, 1986).

Several studies have investigated the relation of codon usage bias with other

biological phenomena such as the abundance of tRNAs (Ikemura, 1985), gene expression levels (Sharp and Li, 1987), as well as gene length and GC composition (Belshaw et al., 2007). Many theories have come up to explain the reasoning behind codon usage bias, including translational selection (Grantham et al., 1981), replicational-transcriptional selection (McInerney, 1998) and mutational bias (Levin and Whittome, 2000). However, the results are so varying that it is difficult to comprehend if these biological phenomena result in codon bias or are a result of codon bias; quite similar to the chicken-and-egg problem (Forsdyke, 2012)!

However, quantification of codon usage bias within the genome of a viral species can give insights into the evolution of genes encoded by overlapping reading frames. Several analyses have been performed to demonstrate that novel genes amongst the set of overlapping viral genes have significantly different codon usage patterns (Pavesi, 2000; 2006; Pavesi et al. 2013). Various open-source programs to calculate the codon usage are now available such as Codon W (Peden, 1999) and Sequence Manipulation Suite (Stothard, 2000).

3.2 Methods

3.2.1 Relative Synonymous Codon Usage (RSCU)

Relative Synonymous Codon Usage (RSCU) values are calculated using the frequencies of each codon type in a nucleotide sequence. The RSCU value for a codon i is defined as:

$RSCU_i$ = *observed(i)/(expected(i),*

where "observed(i)" is the observed frequency of codon i in a gene and "expected(i)" is the frequency expected assuming equal usage of synonymous codons for an amino acid in a gene.

The RSCU index is a measure to assess whether a sequence shows a preference for particular synonymous codons.

3.2.2 Correlation analysis and codon usage

The comparison of the RSCU values obtained from an overlapping gene with that of a non-overlapping gene of the same viral genome was calculated by means of Pearson's correlation coefficient "r" (Sharp and Li, 1987). This correlation coefficient ranges from -1 to +1.

- Concordant Correlation: A value of +1 indicates an identical or concordant relationship,
- Discordant Correlation: A value of -1 indicates a highly dissimilar or discordant relationship.

Studies indicate that newly acquired genes have a tendency towards discordant codon usage, when compared to the rest of the genome. (Pavesi, 2000; 2006; Pavesi et al., 2013)

3.2.3 Datasets

GenBank

GenBank is a comprehensive, open-access genetic sequence database, maintained by the National Center for Biotechnology Information (NCBI) at the

National Institute of Health (NIH) (Benson et al., 2013). It is a part of the International Nucleotide Sequence Database Collaboration (INSDC), the combined effort by GenBank itself, DNA Databank of Japan (DDBJ), and European Molecular Biology Laboratory (EMBL), which ensures uniform and comprehensive collection of sequence information worldwide. They primarily receive nucleotide sequences from the scientific community and data exchange occurs on a daily basis. GenBank can be accessed through the NCBI Entrez retrieval system at ncbi.nih.gov/genbank.

Poorly annotated accessory genes in the GenBank database

SARS-CoV is a relatively new virus and was characterized after its initial discovery in 2003. The sequences deposited in the GenBank (Benson et al., 2013) for its accessory genes/proteins are not fully annotated and there is ambiguity in the name of the gene or protein in several cases. For example, an “orf 9b” (which is the locus tag of the orf9b gene as seen in the SARS-CoV reference sequence NC_004718) search returns just a few results from the SARS-CoV gene/protein database. Some entries of this protein appear under the name “orf 13” as identified by Marra et al. (2003). Even though this protein has been recombinantly produced, its 3D-structure is available, and several functional annotation studies already performed, the protein continues to be named “hypothetical protein sars9b” with inference of “non-experimental evidence, no additional details recorded” in the GenBank database. The same is true for orf3a and orf3b nucleotide sequences. The use of search engines and download parameters fails for these viral sequences. In order to extract all the available orf9b, orf3a, and orf3b nucleotide sequences for this study,

SARS-CoV full-length genomic sequences were collected from GenBank. Similarly, full-length genomic sequences of MNV were collected to extract the overlapping accessory Vf1 gene.

Creation of a local database

1051 full-length nucleotide sequences of SARS-CoV in the GenBank were retrieved. Out of these, only the complete genome sequences (156 in total) were collected and saved in a local database. Similarly, for Murine Norovirus, 28 full-length genome sequences were collected. All the accession numbers are listed in Appendix II. In this study, a HBV reference sequence (NC_003977) was also used for comparison purposes.

3.2.4 Comparison of codon usage of overlapping and non-overlapping gene sets

Codon usage analysis was done using the “sequence analysis program” within the Sequence Manipulation Suite (Stothard, 2000). This program accepts one or more nucleotide sequences and returns the number and frequency of each codon type. RSCU values for each codon were then calculated using these frequencies as explained above.

The comparison was done at the level of codon usage between overlapping and non-overlapping coding regions. The RSCU values obtained from an overlapping gene were then compared with those of the non-overlapping regions of the same viral genome by means of the Pearson correlation coefficient “r”.

The first step in this analysis was to establish a relationship between the codon usage in overlapping and non-overlapping genes of SARS-CoV. The non-overlapping regions of the genome (Fig. 9) were combined into a single unit including the orf1a, orf1b, spike, envelope, membrane, and orf6 genes. On the other hand, the overlapping genes under study here, full-length nucleocapsid, orf9b gene, orf3a and orf3b genes were considered distinct sets of data. The remaining accessory genes (orf7 and orf8) were not included in this comparison, because they contain partial overlaps of very few nucleotides.

Similarly, a relationship between the codon usage in overlapping and non-overlapping genes in the MNV genome were established. The non-overlapping genes of MNV (Fig. 13) were combined into a single unit including orf1 and orf3, and orf2 (VP1) and orf4 (Vf1) genes were considered distinct sets of data. In case of HBV, the non-overlapping third portion of its genome was combined into a single unit, and the overlapping portion of the P gene and the full-length S gene were considered distinct datasets.

3.3 Results

Correlation analysis shows that the overprinting orf9b gene of SARS-CoV has a significant degree of discordance when compared to the rest of the genome (Table 1, next page). It can therefore be concluded that this internal overlapping gene exhibits a choice of synonymous codons, highly different from that occurring in the non-overlapping gene set of SARS-CoV, with an r-value of -0.01.

Virus	Protein	Full-length (amino-acid residues)	Length of overlap	Length of non-overlapping region	Correlation coefficient (r)
SARS-CoV	N	422	98	324	0.62
	Orf9b	98	98	complete overlap	-0.01
SARS-CoV	Orf3a	274	134	140	0.59
	Orf3b	154	134	20	0.24
MNV	VP1	541	213	328	0.60
	Vf1	213	213	complete overlap	-0.08
HBV	P	843	400*	443	0.83
	S	226	226	complete overlap	0.55

TABLE 1 Correlation between the codon-usage patterns of the N and orf9b genes as well as the orf3a and orf3b genes of SARS-CoV, the VP1 and Vf1 genes of MNV, and the overlapping P and S genes of HBV. Also shown is the length of these overlapping proteins along with the length of overlapping and non-overlapping regions in each protein. *It has to be noted that the P gene in HBV is also overlapping with four other HBV genes.

For example, out of the eight proline residues in the orf9b protein, five (63%) are coded by CCC (see Appendix III), whereas in the non-overlapping gene set, less than 9% of proline residues use this codon (Shukla and Hilgenfeld, 2015). On the other hand, concordant relationship is seen between the

overprinted N gene and the non-overlapping gene set of SARS-CoV (r-value of 0.62). In MNV, the overprinting gene Vf1 has also a significantly higher degree of discordance (r-value of -0.08) when compared to its non-overlapping counterparts (r-value of 0.60). However, in the partially overlapping orf3a and orf3b genes of SARS-CoV and in the P and S gene of HBV, the r value is positive for both genes, one being higher than the other.

For the purpose of comparison, codon-usage analysis for other betacoronaviruses containing a hypothetical overlapping "internal" gene within their nucleocapsid gene was also carried out.

Betacorona-virus	**Protein**	**Length (amino-acid residues)**	**Length of overlap**	**Correlation coefficient (r)**
BCoV	Nucleocapsid	448	207	0.67
	internal protein	207	207	-0.11
MHV	Nucleocapsid	455	136	0.66
	internal protein	136	136	0.00
MERS-CoV	Nucleocapsid	411	112	0.58
	hypothetical internal protein	112	112	-0.13

TABLE 2 Correlation between the codon-usage patterns of the overprinted N and its overprinting internal genes of other members of the genus Betacoronavirus i.e. BCoV, MHV and MERS, each with the non-overlapping coding regions in their genome, respectively.

The nucleocapsid genes of Bovine Coronavirus (BCoV), NCBI accession number NC_003045; Mouse Hepatitis Virus (MHV), NCBI accession number AC_000192; and the newly discovered Middle East Respiratory Syndrome Coronavirus (MERS-CoV), NCBI accession number NC_019843, display a similar positive codon usage correlation when compared to the rest of the genome, with r-values of 0.67, 0.66, and 0.57, respectively, whereas their corresponding “internal” genes have r-values of -0.11, 0.00, and -0.13, respectively (Table 2).

3.4 Discussion

Previous work has suggested codon usage as a measure to determine the relative age of a gene (Rancurrel et al., 2009). A discordant relationship in the codon usage of a particular gene, when compared with the rest of the genes, suggests a relatively more recent acquisition of the gene (Pavesi, 1997; Pavesi et al., 2013). The overprinting orf9b gene of SARS-CoV and the Vf1 gene of MNV display discordant codon usage from the rest of their genomes. Hence, it can be concluded that the orf9b gene of SARS-CoV and the Vf1 gene of MNV are novel genes. Both the partially overlapping orf3a and orf3b genes of SARS-CoV display positive correlation with the rest of the genome. However, the degree of concordance is higher in the orf3a gene suggesting that the orf3a gene evolved earlier when compared to the orf3b gene. In HBV too, both the P and S genes display a high degree of positive correlation. Here, it has to be noted that the majority of the genome in HBV is overlapping and the positive correlation of the S gene might be due to the fact that HBV is an old virus and over the course of evolution, both the overlapping genes

have evolved and adapted to accommodate the codon usage of the rest of the genome. These results also affirm the earlier proposal by Belshaw et al. (2007) that in case of internal overlaps, the longer of the two overlapping genes, denoted as “internal primary”, is anscestral relative to the shorter overlapping gene, denoted as “internal secondary” (Belshaw et al., 2007) .

4. Effects of overlaps on the protein products

It has been suggested previously that the overlapping protein products of viral genomes have a tendency to be disordered (Karlin et al., 2003; Rancurrel et al., 2009).

4.1 Intrinsically disordered proteins

"Intrinsically disordered proteins", "flexible proteins", "unstructured proteins" or "natively unfolded proteins" are disorder-related terms which are often interchangeably used for those proteins which lack a well-defined three-dimesnional (3D) structure (Dunker et al., 2001; 2008; Dyson et al., 2005; Gu and Hilser, 2009) . There are different degrees of disorder which can exist, ranging from fully unstructured proteins, random coils, molten globules to large multidomain proteins connected by flexible linkers (Dunker et al., 2001; Dyson et al., 2005; Gu and Hilser, 2009). Most researchers in this field of study use and define the disorder-related terms based on the degree of disorder, localization of disorder (global or region-specific), and several other factors governing their research foci; therefore currently, there is a lack of uniform definition to distinguish between these terms (Gu and Hilser, 2009). An intrinsically disordered protein (IDP) can rapidly and reversibly convert into various structural forms under different thermodynamic conditions. They are also known to adopt ordered structure under certain physiological conditions, such as binding to another macromolecule. This dynamic property of rapidly interconverting between various structural forms allows them to perform intricate functional roles in biological systems (Dunker et al., 2001; 2008; Dyson et al., 2005; Xie et al., 2007).

4.2 Disorder Predictors

A variety of prediction tools have been developed to predict intrinsically disordered protein regions (Romero et al., 1997; Jones and Ward; 2003; Linding et al., 2003a;b; Yang et al., 2005; Gu et al., 2006; Sickmeier et al., 2007). Each algorithm identifies protein disorder within its pre-defined set of parameters. Most sequence-based predictors use the fractional composition and hydropathy of the 20 amino-acid residues. When using a disorder prediction program, one must know what kind of disorder it identifies and which training data set was used for that predictor.

Romero et al. (2001) provided a ranking of the amino-acid residues beginning from disorder-promoting to order-promoting as follows: K, E, D, P, N, S, Q, G, R, T, A, M, H, L, V, Y, I, F, C, W. Such ranking suggests that the amount of flexibility or disorder depends on which residues are used and in what order. In another study, amino acids were clustered into three groups identified from their relative abundance in ordered, disordered, or ambivalent regions (Zhang et al., 2007). The first group, which was mostly associated with disordered regions, contained small and hydrophilic amino acids (K, E, S, G, A). The second group, which was mostly associated with the ordered regions, were hydrophobic residues (M, H, Y, I, F, C, W) . Lastly, the third group, which was found in almost equal frequencies in both ordered and disordered regions, consisted of mainly hydrophilic amino acids (D, T, Q, N, P, R). Apart from few differences, this finding is in reasonable agreement with the ranking provided by Romero et al. (2001). There were other, similar analyses performed by various research groups in the last few years but there was always one or the

other difference in these studies and hence, no general concensus was reached (Wootton and Federhen, 1993; 1996; Gu et al., 2006; Gu and Hilser, 2009).

4.3 Methods

In an overlapping viral gene set, both the “overprinted” or “anscestral” gene and the overprinting or “*de novo*” gene try to alleviate evolutionary constraints imposed on their coding sequence and so higher disorder content at overlapping regions can be seen. It has also been shown that *de-novo* genes have highly unsual sequence composition and greater disorder content when compared with their overprinted counterparts (Rancurrel et al., 2009). Based on these findings, the relative age of the overlapping gene sets can be compared by inspecting the disorder content in the overlapping proteins.

4.3.1 Comparison of disorder content of overlapping protein sets

Disorder predictions were carried out on each set of overlapping proteins, namely the SARS-CoV N and orf9b as well as orf3a and orf3b proteins; the MNV VP1 and Vf1 proteins; and the HBV P and S proteins, to predict their evolutionary age relative to each other. The overlapping protein sequences were retrieved from the previous collection of complete genome sequences of SARS-CoV and MNV (Chapter 3). Accession numbers are provided in Appendix II. The overlapping sequence of the P and S proteins of the HBV reference sequence (NC_003977) were also used for comparison purposes (Table 1, Chapter 3).

To predict the disorder content in overlapping proteins, the DisProt VSL2 intrinsic-disorder prediction program was used. This predictor is a neural network trained on ordered and disordered protein sets and uses attributes such as frequency, flexibility, and hydropathy of each amino-acid residue (Obradovic et al., 2003). Each amino-acid residue in the protein sequence is assigned a value between 0 and 1, which measures the intrinsic disorder; 0 being highly ordered, hence, contributing to a well-defined, ordered 3D structure, and 1 being highly disordered. The results are then plotted in a graph for each set of overlapping proteins under study.

4.4 Results

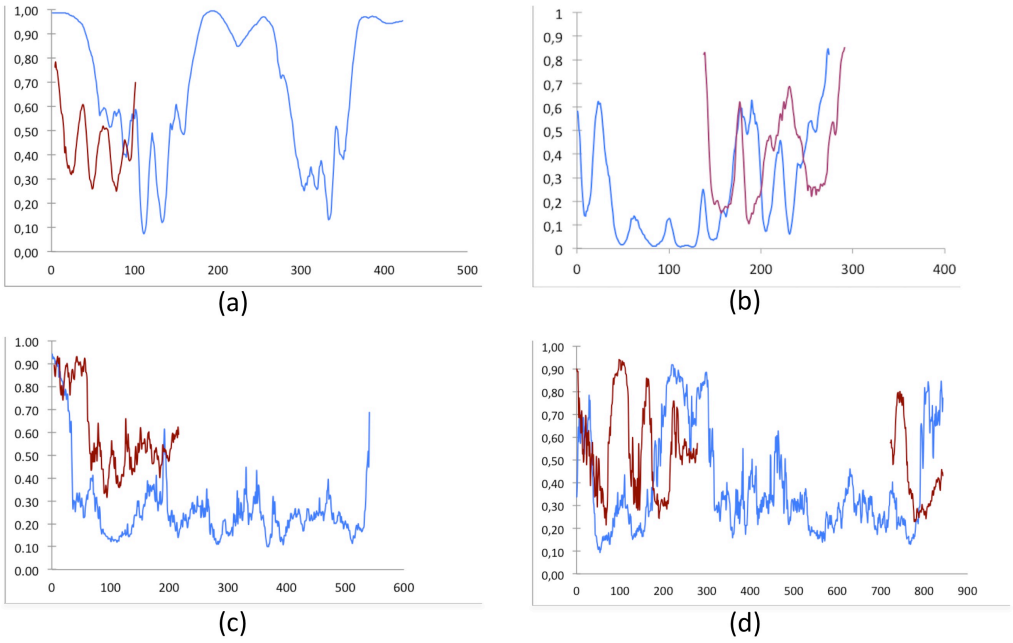


FIGURE 16 Graphical illustration of the disorder content in the overlapping protein set of (a) SARS-CoV N (blue) and orf9b (red), (b) SARS-CoV orf3a (blue) and orf3b (red), (c) MNV VP1 (blue) and Vf1 (red), and (d) HBV P (blue) and S (red). The X-axis represents the amino-acid residues and the Y-axis represents the degree of disorder as calculated by Disprot VSL2.

Figure 16 shows a graphical representation of the disorder content in the overlapping protein set of SARS-CoV N and orf9b, SARS-CoV orf3a and orf3b, MNV VP1 and Vf1, and HBV P and S. The disorder content for each amino-acid residue lies between the range 0 and 1, 0 being highly ordered and 1 being highly disordered. The threshold between order-disorder is 0.5. It can be inferred from the graphs above that the protein regions containing overlaps are predicted to have higher disorder content when compared to the non-overlapping regions. The percentage disorder content in the overlapping and non-overlapping regions for the same set overlapping proteins are depicted in the table below (Table 3).

Virus	**Proteins**	**% disorder (overall)**	**% disorder (overlapping region)**	**% disorder (non- overlapping region)**
SARS-CoV	N	72.7%	90.8%	67.2%
	Orf9b	32.6%	32.6%	complete overlap
SARS-CoV	Orf3a	19.3%	29.3%	8.9%
	Orf3b	28.5%	22.4%	70%
MNV	VP1	7.4%	16.4%	1.5%
	Vf1	65.7%	65.7%	complete overlap
HBV	P	27.3%	44.5%	11.7%
	S	52.5%	52.5%	complete overlap

TABLE 3 Disorder content in the overlapping proteins N and 9b, orf3a and orf3b of SARS-CoV, VP1 and Vf1 of MNV, and P and S of HBV calculated by using the DisProt VSL2 program.

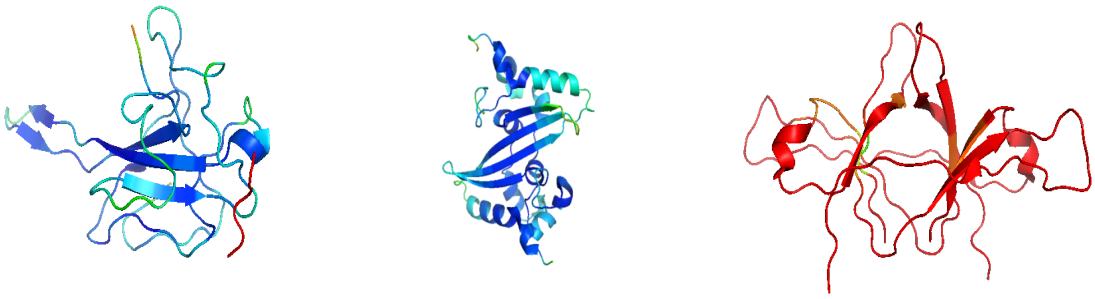
If the overlapping proteins set consists of an anscestral protein and a relatively novel protein, the *de novo* protein tends to have a higher degree of disorder (Rancurrel et al., 2009). In Murine norovirus, this result is quite clearly illustrated. The newly acquired Vf1 accessory protein, (McFadden et al., 2011) has a much higher percentage of predicted disorder content (65.7%) than the VP1 structural protein (7.4% overall). Moreover, the VP1 structural protein has an even smaller predicted disorder content in its non-overlapping part (1.5%).

In the overlapping polymerase and surface protein of HBV, the predicted disorder content in the entire polymerase protein is 27.3%, less than the disorder content in its overlapping region (52.5%).

In case of SARS-CoV, a similar trend can be observed in the orf3a protein (Fig. 16, Table 3). But the non-overlapping C-terminal region of 20 amino-acid residues in the orf3b protein (full-length: 154 amino-acid residues) is predicted to be highly disordered. It is also interesting to observe the structural protein N of SARS-CoV, which is predicted to have a high disorder content (72.7%), in contrast to the orf9b protein predicted (predicted disorder content 32.6%). This result is quite opposite to the expected!

Fortunately, X-ray crystallographic structures are available for this overlapping protein set (Fig. 17), (Meier et al., 2006; Yu et al. 2006; Saikatendu et al., 2007). There are two crystal structures of the SARS-CoV nucleocapsid protein, one for the N-terminal domain (NTD) [PDB id: 2OFZ] (Saikatendu et al., 2007), and the other for the C-terminal domain (CTD)

[PDB id: 2GIB] (Yu et al. 2006); and one crystal structure for the orf9b protein [PDB id: 2CME] (Meier et al., 2006).



(a) (b) (c)

FIGURE 17 Structure of SARS-CoV proteins, colored according to the B-factor: (a) NTD of nucleocapsid protein with average B factor 11.2 Å^2 [2OFZ] (Saikatendu et al., 2007), (b) CTD of nucleocapsid protein with average B factor 17.9 Å^2 [2GIB] (Yu et al. 2006), and (c) dimer of the orf9b protein with average B factor 100.8 Å^2 [2CME] (Meier et al., 2006).

In order to explain the unexpected overall disorder prediction result for the SARS-CoV nucleocapsid and the orf9b proteins, these available crystal structures were compared according to their average atomic temperature factor (B-factor).

A number of factors contribute to the value of the B-factor, namely the degree of disorder of the crystal, rigid-body movements of the molecules, experimental errors, etc., but in principle, it can be said that the lower the B-factor, the better ordered is that protein domain.

PDB ID	Protein	Average B Factor
2OFZ	SARS-CoV NTD	11.2 $Å^2$
2GIB	SARS-CoV CTD	7.9 $Å^2$
2CME	SARS-CoV orf9b	100.8 $Å^2$

Table 4 B-factors for the crystal structures of SARS-CoV Nucleocapsid NTD and CTD; and for the SARS-CoV orf9b lipid-binding protein.

After inspecting the 3D structures, it is beyond doubt that the crystallized NTD (residues 47-175) and CTD (residues 270-370) of the nucleocapsid protein have a well-defined 3D structure, with very low average B factors (Table 4). In contrast, the orf9b protein is rather flexible with a much higher average B-factor (Meier et al., 2006). However, it has to be noted that almost 50% of the overlapping part of the nucleocapsid protein (residues 1-46) was excluded from the crystallized NTD fragment, as they are believed to be diordered on the basis of secondary structure prediction, limited proteolysis experiments, and sequence conservation. In the following section, the discrepancy between the result obtained by inspecting the crystal structures and disorder prediction (in case of the overlapping part of the SARS-CoV N and orf9b proteins) will be discussed.

4.5 Discussion

Disorder prediction results suggest that since there is less overall disorder observed in the overprinted proteins, VP1 of MNV and P of HBV evolved prior

to their overprinting counterparts. The disorder content in both the proteins (VP1 of MNV and P of HBV) is predicted to be higher in the overlapping region and substantially lower in the non-overlapping regions. This result supports the hypotheses developed by Rancurrel et al. (2009) on overlapping proteins of various virus families. This trend is also demonstrated by their respective overprinting counterparts, i.e. the Vf1 protein of MNV and the large S protein of HBV. The set of partially overlapping accessory proteins of SARS-CoV, namely orf3a and orf3b, also follows the same trend but in this case, the differences between the overall disorder content between these proteins are very small (overall disorder content: 28.5% for the orf3b protein, and 19.3% for the orf3a protein). Based on this data, it is difficult to predict the relative ancestry of this protein set.

Interestingly, the result of disorder prediction for the SARS-CoV N protein and its overprinting counterpart orf9b gave an altogether different perspective. When comparing just the overlapping part of the SARS-CoV nucleocapsid NTD and the orf9b protein, it is observed that the first 46 N-terminal residues of the NTD were excluded from the fragment that was crystallized. The well-defined NTD (residues 47-175) comprises an antiparallel β-sheet core and protruding from it is a β-hairpin (Fig. 17a). The average B-factor for the crystallized fragment of NTD is 11.2 $Å^2$, thereby revealing reduced flexibility (Saikatendu et al., 2007). The orf9b protein forms a two-fold symmetric dimer comprising two adjacent β-sheets (Fig. 17c). Electron density for a lipid molecule was detected in the central hydrohobic cavity between the two monomers (Meier et al., 2006).

The orf9b protein appears to be much more flexible than the NTD, with an average B-factor of 100.8 $Å^2$ (Meier et al., 2006). This huge difference in the B values between the two overlapping protein regions shows that the orf9b polypeptide chain is very flexible. Also, two segments between residues 1 - 8 and 26 - 37 in the orf9b protein were not visible in the electron-density maps. In contrast, in the NTD crystal structure of the N protein, all residues are well-defined by electron density (Saikatendu et al., 2007).

The full-length SARS-CoV N protein comprises two well-defined domains (NTD and CTD), and three inherently disordered regions (IDR's), between residues 1-46, 176-269, and 371-422. These IDR's are required for proper biological functioning of the SARS-CoV N protein (Chang et al., 2014), but in turn contribute to the high overall disorder content. Comparison of the X-ray crystallographic structure of both the proteins showed that the N protein is divided into two well-defined domains unlike the more flexible orf9b protein (Meier et al., 2006; Yu et al. 2006; Saikatendu et al., 2007). Thus, only by inspecting the crystal structures of these two overlapping protein segments, it can be concluded that the orf9b protein has evolved recently. However, the disorder prediction results on this protein set would have failed to support the same.

5. Prediction of RNA structure at the sites of initiation of alternative reading frames

5.1 Molecular mechanisms for the initiation of the alternative reading frame in overlapping genes

In RNA viruses, overlapping proteins can be produced via a number of non-canonical translation mechanisms. These mechanisms can be classified into two broad categories: a) non-canonical translation initiation, and b) non-canonical translation elongation and termination (Firth and Brierely, 2012). While the former requires the presence of a secondary initiator codon on its messengerRNA (mRNA), the latter does not necessarily require another AUG. Generally speaking, there are two common mechanisms used in the translational initiation of downstream open reading frames of multicistronic mRNA: Leaky scanning of ribosomes (Kozak, 1986, 1989), and internal ribosomal entry (Thiel and Siddell, 1994). Moreover, there are two common mechanisms used in non-canonical translation initiation: programmed ribosomal frameshifting (Jacks et al., 1988; Brierley et al., 1989; Brierley and Dos Ramos, 2006; Dinman, 2012), and stop-codon readthrough (Beier, 1984; Honigman, 1991; Orlova, 2003).

The dual-coding, subgenomic (sg) mRNAs of the SARS-CoV coding for the N and the orf9b proteins, and for the orf3a and orf3b proteins, both contain a secondary start codon (AUG). The same is the case in the dual-coding sg-mRNA of MNV coding for the VP1 and Vf1 proteins, and in the polycistronic sg-mRNA coding for the HBV S protein. Hence, these viruses should make use of the non-canonical translation initiation mechanism (described in details

below) for the creation of their overprinting proteins: SARS-CoV orf9b, SARS-CoV orf3b, Vf1 of MNV, and S of HBV.

5.1.1 Leaky Scanning

As a general rule, translation initiation occurs exclusively at the first AUG codon at the 5' end of the mRNA (Kozak 1983, 1999, 2002). However, it has been demonstrated in dual-coding mRNAs that if a second AUG codon lies in close vicinity, the small ribosomal subunit might flutter back and forth between these two AUGs, sometimes bypassing its attachment to the first AUG, so that the translation of the downstream ORF begins (Kozak, 1986) .

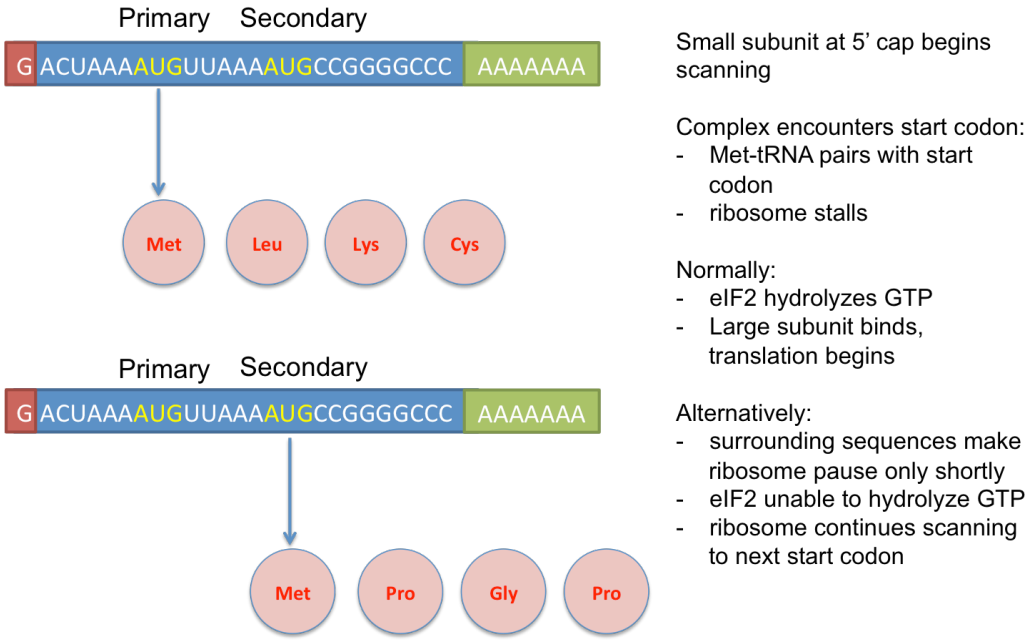


FIGURE 18 Mechanism of leaky ribosomal scanning (adapted from Watkiss, 2010).

This fluttering of the small ribosomal subunit is called leaky scanning, which has been shown to depend on a short consensus nucleotide sequence around the AUG initiator codon known as Kozak sequence (Kozak, 1986). Mutagenesis experiments have identified this consensus sequence around the AUG initiator codon as GCCRCC**AUG**G (Kozak, 1986).

The numbering of the mRNA sequence begins with the initiator codon AUG, with base A numbered as position 1 and the preceding base numbered as -1 (Kozak, 1986). It has been shown that positions -3 and +4 are most important in determining the efficiency of translation initiation at that AUG codon. The -3 position could be any purine; however, adenine is preferred over guanine. In an optimal Kozak sequence, guanine should be present at the +4 position (Kozak, 1997).

A 10-fold increase in the translational efficiency has been demonstrated in the presence of an optimal Kozak sequence as opposed to a weak Kozak sequence (Kozak, 1986). If the first AUG codon in a dual-coding mRNA is related to a weak or sub-optimal Kozak sequence, then the ribosome could bypass the first AUG and start translation at the second AUG, irrespective of the context (strong or weak) of the Kozak sequence of the second downstream AUG (Kozak, 2002).

If the second AUG is in a different reading frame than the first AUG, an entirely different protein is translated; and, if the second AUG is in the same reading frame as the first AUG, an N-terminally truncated version of the same protein is created. The leaky scanning mechanism fails if the second initiation codon lies far away from the 5’ end of the overprinted ORF. It has been

speculated that the large ribosomal subunit can mask the downstream AUG (Kozak, 1995).

Many viruses, especially +ssRNA plant viruses, use the leaky scanning mechanism to produce different functional proteins from the same gene (Dreher and Miller, 2006; Ryabova et al., 2006; Castano et al., 2009; Watkiss, 2010). In case of the orf9b gene of SARS-CoV and of the Vf1 gene of MNV, the translation initiation site lies less than 15 nucleotides away from the 5' end of the respective overprinted ORF. Both these proteins have been shown to be a product of leaky ribosomal scanning (Xu et al., 2009; Mc Fadden et al., 2011).

5.1.2 Internal ribosomal entry

Unlike leaky scanning, the use of an internal ribosomal entry site is independent of the first start codon in an open reading frame (Jackson and Kaminsky, 1995). An internal ribosome entry site (IRES) is a conserved nucleotide sequence that allows for the initiation of translation at a start codon present in the middle of a messenger RNA (mRNA). Various secondary and tertiary structure elements contribute to the effectiveness of an IRES. They are thought to bind directly or indirectly to the components of the translational machinery (Kieft, 2008). Many viruses employ this mechanism for the translation of its multicistronic mRNA (Jackson et al., 1990; Brown et al. 1992; Liu and Inglis, 1992; Thiel and Siddell, 1994; Rota et al., 2003). However, no consensus sequence or secondary structure elements are reported till date and the structural elements vary from virus to virus.

It is hence seen that the sequence flanking the initiation codon of an open reading frame and base pair interactions within this sequence are important in determining the expression of that gene. The presence of secondary or tertiary structure elements in the mRNA influences its translational efficiency (Kozak, 2005; Marzi et al., 2007; Kieft, 2008). In this work, I investigate the presence of consensus secondary and tertiary RNA structure elements at the site of frameshift, which might assist in one of the two molecular mechanism explained above for the initiation of the alternative reading frame in overlapping genes. The secondary as well as tertiary structure of RNA have been predicted and analyzed at the site of frameshift, to throw some light onto this question.

5.2 Methods

Prediction of RNA secondary and tertiary structural elements at the site of frameshift in the overlapping orf3a/orf3b genes as well as the N/orf9b genes of SARS-CoV, orf2/orf4 of MNV, and the P/S genes of HBV was carried out using the RNAfold web server within the Vienna RNA package 2.0.0 (Lorenz et al., 2011). Three-dimensional RNA structure prediction was also carried out using RNAComposer (Popenda et al., 2012).

5.2.1 Vienna RNA package

The ViennaRNA Package 2.0.0 (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi, last accessed on 16.09.13) contains several stand-alone programs for the prediction and comparison of RNA secondary structures. For the present analysis, the RNAfold web server within this package was used. RNAfold predicts secondary structures of single-stranded RNA or DNA sequences

based on minimum Free Energy calculations. The current RNA size limit is 10,000 nucleotides (Lorenz et al., 2011).

5.2.2 RNAComposer

RNAComposer is a recently developed automated tool for 3D RNA structure prediction. The input to RNAComposer is a user-defined secondary structure, in the present study obtained from the RNAfold server. This method is based on the RNA FRABASE database and converts the RNA secondary structure input to its corresponding tertiary structure elements. The RNA FRABASE database is derived from 2270 RNA structures deposited in the PDB (Popenda et al., 2012). Below is an example of a typical dot bracket input to RNAComposer:

SARS coronavirus, complete genome
NCBI Reference Sequence: NC_004718.3
Showing 41bp region from base 28110 to 28150 (28130)
GenBank Graphics

```
>gi|30271926:28110-28150 SARS coronavirus, complete
genome

ACAAATTAAAATGTCTGATAATGGACCCCAATCAAACCAAC

ACAAAUUAAAAUGUCUGAUAAUGGACCCCAAUCAAACCAAC

.............((.((((..(((...)))))))).))....
```

5.2.3 Computer prediction of RNA secondary and tertiary structural elements

RNA secondary and tertiary structure predictions were carried out for surrounding nucleotides flanking the initiation site of the alternative reading frame in the overlapping gene sets in order to probe the molecular

mechanism behind the frameshift. Sequences were collected at the site of initiation of the alternative reading frame for the SARS-CoV orf9b and orf3b genes, the MNV Vf1 gene, and the HBV S gene. The start co-ordinates are indicated in Figures 11, 13, 14, and 15, respectively. Five window frames (described below) were selected for each overlapping gene to make sure that the obtained secondary structure element is conserved and not just arbitrary.

20-nt window: 10 nt….begin of overlap……10 nt = 21-nt RNA sequence

40-nt window: 20 nt….begin of overlap……20 nt = 41-nt RNA sequence

60-nt window: 30 nt….begin of overlap……30 nt = 61-nt RNA sequence

80-nt window: 40 nt….begin of overlap……40 nt = 81-nt RNA sequence

100-nt window: 50 nt….begin of overlap…..50 nt = 101-nt RNA sequence

For each overlapping gene, 5 sets of sequences were saved in FASTA format. Multiple sequence alignment was performed to obtain a consensus sequence using ClustalX version 2.1 (Larkin et al., 2007). These 5 sequences were then used as an input to the RNAfold webserver to predict the secondary structure. Since the secondary structures obtained were the same amongst the 5 window sets, the 40-nt window output was selected for further analysis. The output for all the 4 overlapping genes was also saved in dot bracket format. This output served as an input to RNAComposer.

RNAComposer predicted the 3D structure of RNA based on the secondary structure predictions of RNAfold and the RNA FRABASE database.

5.3 Results

The predicted RNA secondary structures at the site of the translational initiation in the overprinting genes orf9b of SARS-CoV, Vf1 of MNV and S of HBV are shown below. No nonsensus RNA-secondary structure elements can be seen at the translation initiation site of these overprinting proteins under study.

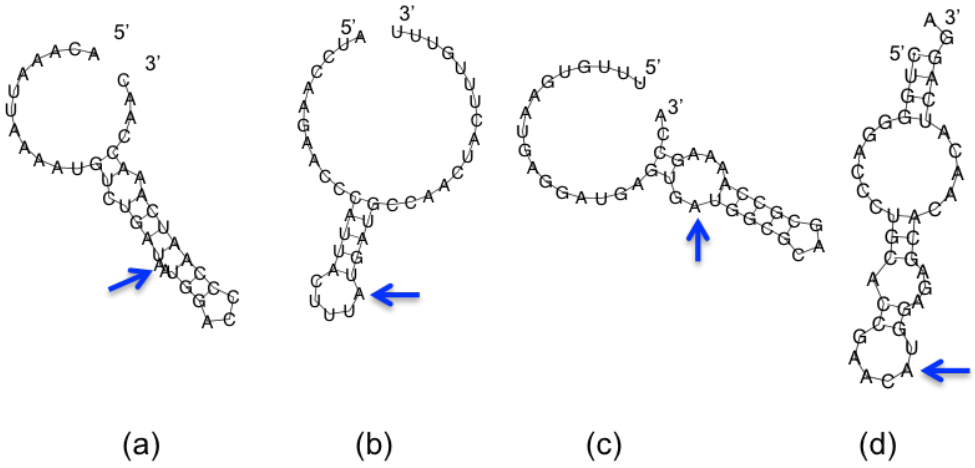


FIGURE 19 Predicted RNA secondary structures of the 40 nucleotides flanking the site of translational initiation in the overlapping (a) orf9b gene of SARS-CoV, (b) orf3b gene of SARS-CoV, (c) Vf1 gene of MNV, and (d) S gene of HBV, using the RNAfold web server within the Vienna RNA package 2.0.0 (Lorenz et al., 2011) . The arrow shows the intiation site of alternative reading frames.

The predicted RNA tertiary structures at the site of the translational initiation in the overprinting genes orf9b of SARS-CoV, Vf1 of MNV and S of HBV are also shown in the following page.

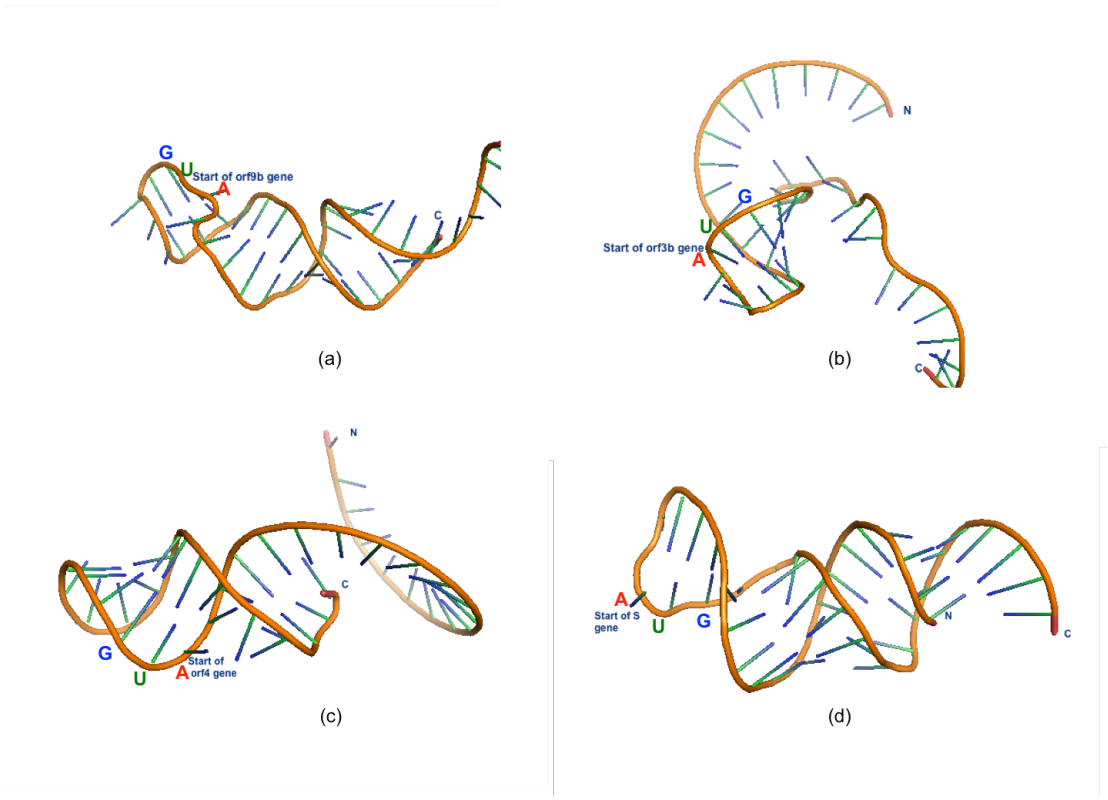


FIGURE 20 Predicted RNA 3D structure of the 40 nucleotides flanking the site of translational initiation in the overlapping (a) orf9b gene of SARS-CoV, (b) orf3b gene of SARS-CoV, (c) Vf1 gene of MNV, and (d) S gene of HBV, using the RNAComposer web server.

5.4 Discussion

As seen from the results, it can be concluded that there is no concensus RNA secondary or tertiary structure element present downstream or upstream in the overprinting genes of SARS-CoV, MNV, and HBV. However, when looking at the results individually, it can be hypothesized that for the translation of the

orf9b sub-genomic mRNA via leaky scanning (Xu et al., 2009), the presence of an A-bulged loop one nucleotide upstream from the frameshift site may be of assistance (Figs. 19a, 20a). Point mutations at this site might provide concrete insight into this +1 frameshift. The orf3b gene of SARS-CoV is speculated to use an IRES for its translation initiation as its initiator codon is at the 13th position in the sub-genomic mRNA3 (Snijder et al., 2003). On the other hand, in MNV, due to the presence of a weak Kozak context for the AUG of the VP1 gene (U at -3 and A and +4) , Vf1 is believed to be translated via leaky scanning (McFadden et al., 2011). Figures 19 (b,c) and 20 (b,c) show the presence of the intiator AUG of the alternate reading frame for orf3b of SARS-CoV and Vf1 of of MNV within a stem loop structure. Secondary and tertiary structure elements, including stem loops, are known to slow down the ribosome, favoring ribosomal slippage or allowing enough time for the ribosome to attach to the initiator codon (Jacks et al., 1988; Kozak, 2005) . These elements could also function as internal ribosomal entry sites (Jackson, 2000). It should also be noted that the region upstream of the orf3b initiation codon shows the presence of a consencus UUACUUU sequence in all the orf3b genes, which is responsible for this stem loop formation. The molecular mechanism behind the translation of the S gene in HBV is not fully understood, though a leaky scanning mechanism has been proposed (Zajakina et al., 2004). Figure 19(d) and 20(d), shows the presence of the S gene (HBV) initiation codon within a loop-stem-loop structure which could assist in the translatory mechanism.

6. Mutational model for the evolution of SARS-CoV overlapping accessory protein 9b

6.1 Introduction

In terms of evolution, overlapping genes are considered a mechanism for creating novel proteins (Krakauer, 2000; Rancurrel et al., 2009). However, any point mutation occurring in an overlapping gene region affects two (or more) protein products at the same time. Studies revealing differential selective pressure during the evolution of overlapping viral genes (Miyata and Yasunaga, 1978; Mizokami et al., 1997; Pavesi, 2000; 2006; Jordan et al., 2000; Fujii et al. 2001; Zaaijer et al., 2007) led me to probe the selection pressure acting on the overlapping N and orf9b genes. In this chapter, I will present the analysis of the evolution of orf9b in concert with orf9a (i.e., the N gene) using sequence data of betacoronavirus lineage b.

6.2 Methods

6.2.1 Dataset

70 full-length genomic sequences including one reference sequence of SARS-CoV were retrieved from the locally created databse (Chapter 3). Among these, 37 isolates were from human SARS-CoV, 15 from civet SARS-CoV, and 18 (including the newly discovered SL-CoV-WIV1 (Ge et al., 2013)) from bat betacoronaviruses of clade b. Accession numbers are given in Appendix II. These full-length genomic sequences were parsed and

corresponding gene and amino-acid sequences were collected in a local database for further analysis.

6.2.2 Mutation rate analysis

Mutation rate analysis was performed by first aligning both the nucleotide and protein sequences using ClustalX version 2.1 (Larkin et al., 2007). Redundant sequences (from multiple human patients) were manually removed. DnaSP 5.10 (Librado and Rozas, 2009) was used to calculate the number of synonymous and non-synonymous substitutions in the overlapping gene regions of the N gene and the orf9b gene.

6.2.3 Entropy-plotting

Entropy

Entropy is defined as a measure of uncertainty at each position in a set of aligned nucleotide or protein sequences (Hall, 1999). The cumulative entropy is the sum of all the entropy values calculated at each position in a sequence. For calculating the entropy, sequences are treated as a matrix of characters and the maximum number of different characters found in a column (column of aligned sets of nucleotide or amino-acid sequences) defines the maximum total uncertainty or the “entropy” (Hall, 1999).

The entropy H is calculated by: $H(l) = -\sum f(b,l) \ln(f(b,l)),$

where H(l) is the uncertainty, also called entropy, at position l, b represents a residue type (out of the allowed choices for the sequence under investigation), and f(b,l) is the frequency at which residue b is found at position l.

The frequency of substitutions at each codon site in the overlapping region of the nucleocapsid/orf9b gene was calculated to determine the evolutionary strategy followed by this set of overlapping genes in SARS-CoV. There are 98 codons in the overlapping region, and therefore, the variation of nucleotides in 294 codon positions and their corresponding amino acids in the overlapping nucleocapsid and orf9b gene region of 70 SARS-CoV genomes was measured.

Entropy-plotting of alignments

Entropy-plotting of alignments was carried to determine the variations occurring in the overprinted nucleocapsid (N) protein and the overprinting orf9b protein. In this overlapping region of the SARS-CoV genome, the first nucleotide position of an N codon corresponds to the third nucleotide position of an orf9b codon (N1/9b3), the second position in an N codon corresponds to the first nucleotide position in an orf9b codon (N2/9b1), and the third position in an N codon corresponds to the second nucleotide position in an orf9b codon (N3/9b2) (Fig. 21, Table 5). Hence, variation in the three sets of nucleotide sites, comprising 98 sites each of positions N1/9b3, N2/9b1, and N3/9b2, and the variation in the corresponding 98 amino-acid residues in the overlapping proteins, were studied by plotting the entropy (variability) of the aligned overlapping nucleotide and protein sequences, as implemented in the BioEdit software v7.0.0. (Hall, 1999).

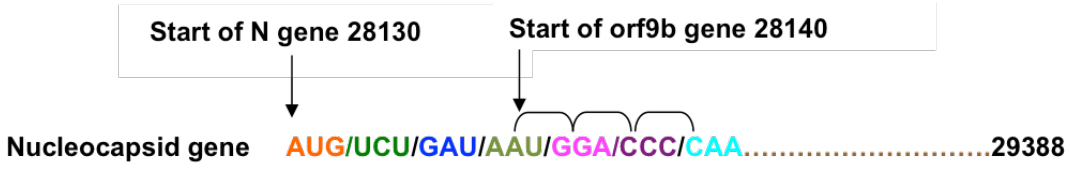


FIGURE 21 The 5' ends of the SARS-CoV N and orf9b genes. At the 10th nucleotide position in the N gene, a +1 frameshift leads to the translation of the overlapping orf9b gene.

N	1	2	3	1	2	3	1	2
9b	3	1	2	3	1	2	3	1
Codon sites	N1/9b3	N2/9b1	N3/9b2	N1/9b3	N2/9b1	N3/9b2	N1/9b3	N2/9b1

TABLE 5 Codon-site substitutions in the two genes. Three types of substitution have to be distinguished: N2/9b1, N3/9b2, N1/9b3.

6.3 Results

6.3.1 The effect of the overlap on the mutation rate in the N and orf9b genes

The ratio ω of nonsynonymous (K_a) to synonymous (K_s) nucleotide-substitution rates is an indicator of selective pressures on genes. A ratio significantly greater than 1 indicates positive selective pressure. A ratio around 1 indicates either neutral evolution at the protein level or an averaging of sites under positive and negative selective pressure. A ratio less than 1 indicates pressure to conserve protein sequence, i.e. “purifying selection” (Hurst, 2002).

Gene regions of:	K_a	K_s	$K_a/K_s = \omega$
Nucleocapsid (overlapping part)	0.41	0.73	0.56
Nucleocapsid (non-overlapping part)	0.37	0.59	0.63
Orf9b	0.53	0.43	1.23

TABLE 6 Synonymous and non-synonymous substitutions in overlapping and non-overlapping regions of the SARS-CoV nucleocapsid gene and in the orf9b gene.

The overlapping gene regions of nucleocapsid and orf9b show differences in the evolutionary rates. The orf9b gene has a K_a/K_s ratio greater than 1 (Table 6), indicating that it is subject to positive selection pressure and is evolving at a faster rate. On the other hand, the overprinted region of the nucleocapsid gene has a K_a/K_s ratio of 0.56, which means that this protein is rather conserved (the K_a/K_s ratio is 0.63 for the non-overlapping part of the N gene). Remarkably, due to the frameshift, the same stretch of genome has thus a different evolutionary rate when coding for each of the two different proteins. This observation prompted further analysis of the nucleotide variations at each of the three nucleotide positions of the codons.

6.3.2 Evolutionary strategy adopted by the N and orf9b overlapping gene set

Upon a point mutation, the position at which nucleotide substitution occurs within a codon reflects whether the substitution would be synonymous or not. Due to the partial degeneracy of codons (for detailed explanation, see

Chapter 1), the nucleotide substitution in the first position of a nucleocapsid codon (N1/9b3) is likely to cause an amino-acid change in N but not in orf9b, whereas substitutions in the third position of an N codon (N3/9b2) are probably non-synonymous in orf9b, but synonymous in N.

The nucleotide variation at the 98 N1/9b3, N2/9b1, and N3/9b2 positions (Fig. 21, Table 5) of the overlapping N/orf9b gene region for all 70 sequences is shown in Fig. 22. The value of cumulative mutational frequency ($\sum(H)$); see Methods) of the overlapping region of the nucleocapsid protein is 4.7 and that of the orf9b protein is 15.3. This difference in the frequency of mutation was somewhat expected, based on the different ω values for the two genes (Table 6). Moving on to the nucleotide level, cumulative entropy values of 3.16, 3.49, and 5.44 for the N1/9b3, N2/9b1, and N3/9b2 codon positions, respectively were obtained. The graphs were calibrated in the range of 0 - 1 for accurate comparison of the results of protein sequences with nucleotide sequences (Fig. 22, next page).

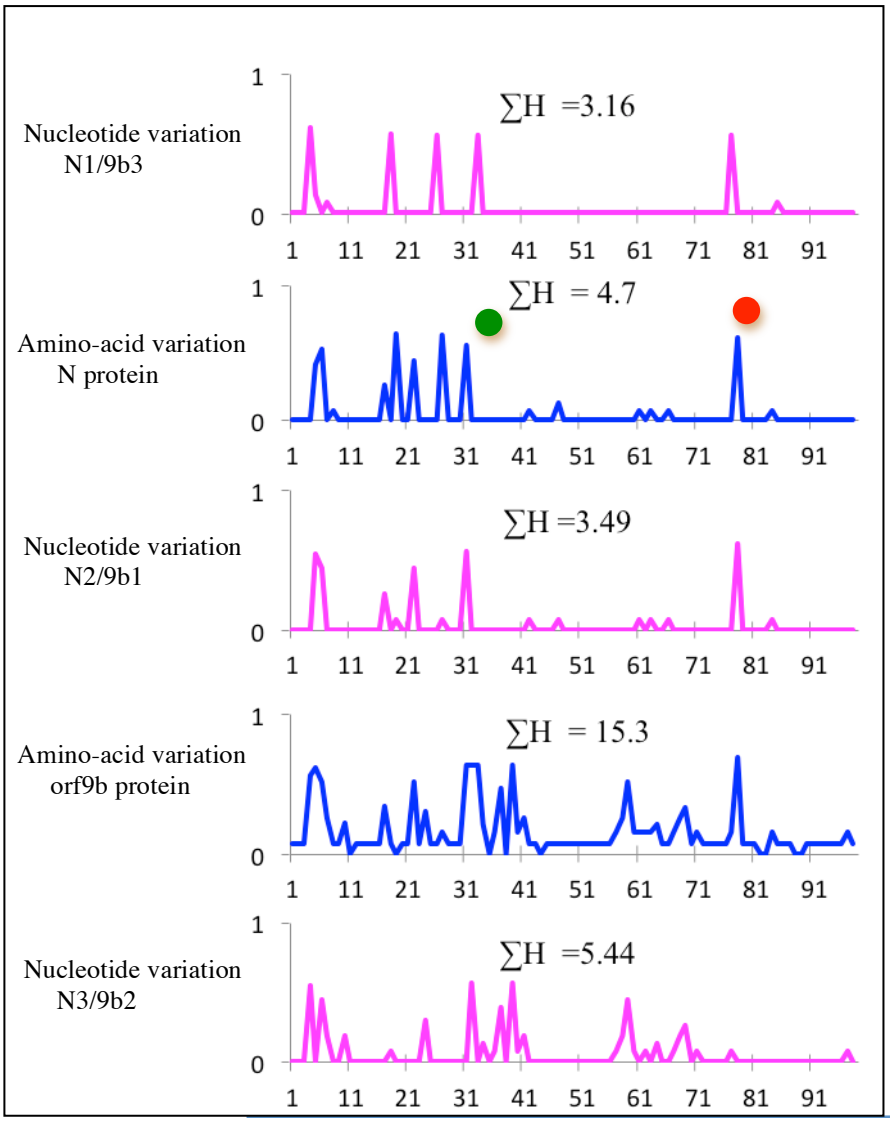


FIGURE 22 Variation of three sets of nucleotides (in magenta): N1/9b3, N2/9b1, and N3/9b2, in relation to the amino-acid variations (in blue) in the genes coding for the overlapping nucleocapsid and orf9b proteins. The x-axis represents the codon sites in case of graphs 1, 3, and 5, i.e. nucleotide variations, whereas in case of graphs 2 and 4, the x-axis represents the amino-acid residue number. Note that the N protein overlaps with orf9b between its residues 4 and 101; however, in graph 2, which represents the amino-acid variations in the N protein, the x-axis is calibrated from 1 to 98 in order to facilitate the comparison with orf9b. The y-axis represents entropy. The green dot indicates the one case of synonymous N1/9b3 substitution that does not lead to an amino-acid exchange in the N protein because of the partial degeneration of the first nucleotide position in a codon (AGA and CGA both code for Arg). The red dot indicates a case of a

two-nucleotide difference as a result of an N1/9b3 and an N2/9b1 substitution that leads to an amino-acid exchange in the N protein. All bat betacoronaviruses of clade b (with the exception of SL-CoV WIV1 (Ge et al., 2013)) have Lys at this position, whereas all civet and human SARS-CoV isolates as well as bat SL-CoV WIV1 have Pro.

The higher rate of amino-acid variation in the orf9b protein is largely determined by nucleotide substitutions at the N3/9b2 sites. All the N3/9b2 nucleotide variations translated to amino-acid changes in the orf9b protein but were silent in the nucleocapsid protein. Amino-acid variations in the nucleocapsid protein were determined by substitutions at the N1/9b3 nucleotides. In one instance, an N1/9b3 nucleotide variation resulted in a synonymous mutation in the nucleocapsid protein (Fig. 22, green dot). The amino acid at this position (nucleotide position 28172) is arginine and this phenomenon occurs due to partial degeneracy of the first nucleotide position (as explained in Chapter 1).

There were a few N2/9b1 mutations that imposed amino-acid variations in both the proteins. An interesting variation, corresponding to a concomitant N1/9b3 and N2/9b1 exchange results in an amino-acid difference at position 81 of the nucleocapsid protein, within its well-ordered and overprinted part. All known genomic sequences of bat betacoronaviruses of clade b feature the AAA triplet (coding for Lys) here, whereas all isolates of civet and human SARS-CoV have CCA (coding for Pro). The exception among the bat beta-CoVs of clade b is the newly discovered SL-CoV WIV1, which is proposed to be the likely originator of SARS-CoV (Ge et al., 2013); the N gene of this virus also has CCA coding for Pro at this position. Thus, there is a two-nucleotide difference between the codons in the bat CoVs (except SL-CoV WIV1) on the

one hand and human or civet SARS-CoV on the other (see red dot in Fig. 22). In the orf9b protein, the corresponding codon (shifted by +1 in frame) is AAG (coding for Lys) in the bat betacoronaviruses of clade b, and CAG (coding for Gln) in the civet and human SARS-CoV sequences as well as in SL-CoV WIV1 (Ge et al., 2013). Thus, only the N2/9b1 variation resulted in an amino-acid change in the orf9b protein and the N1/9b3 nucleotide variation was silent.

6.3.3 Effect of mutations on the three-dimensional structures of the overlapping proteins

Multiple sequence alignment results were parsed to identify amino-acid variations in the overlapping nucleocapsid and orf9b protein sequences. It was found that the majority of mutations occurred in the disordered regions of the N protein. In the overprinted part N-terminal domain (NTD) of the N-protein, mutations were mostly observed in the disordered segment between the amino-acid residues 1 and 46 (Ch 4, section 4.6). Other than that, one prominent mutation, identified at position 81 in the overprinted part of the NTD was observed. In the three-dimensional structure of the SARS-CoV nucleocapsid protein, this Pro residue occupies position 2 of a surface-exposed type-III β turn of the sequence Gly-Pro-Asp-Asp (Saikatendu et al., 2007). All mutations in the SARS-CoV orf9b protein structure fall into the disordered regions (Fig. 23) and hence, these proteins follow the general principle that mutations are more commonly localized in regions of no regular secondary structure.

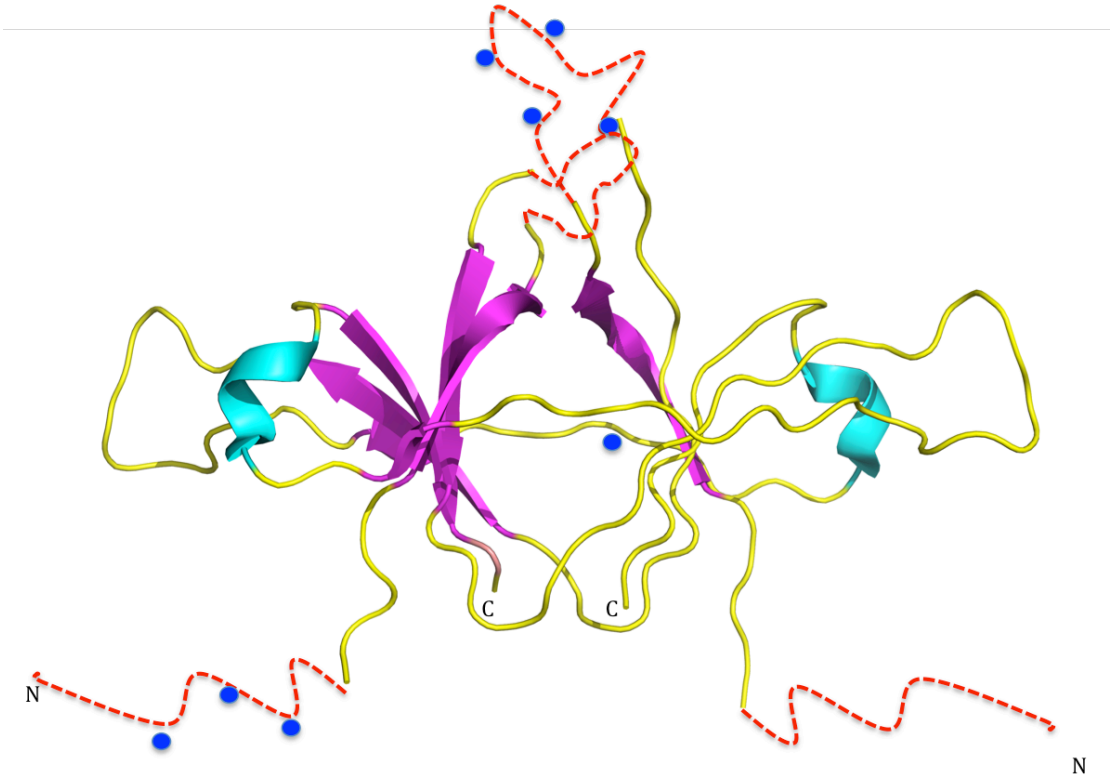


FIGURE 23 Structure of the SARS-CoV orf9b protein dimer, colored according to the secondary structure [2CME] (Meier et al., 2006). α-helices are colored in aqua-blue, β-strands are colored in magenta and the defined non-α, non-β structure are colored in yellow, disoredered loops are colored in red. Blue dots depict the location of the mutations in a monomer.

6.4 Discussion

Extra evolutionary constraints are imposed on overlapping, frame-shifted genes. In a number of overlapping viral genes, a slower rate of evolution has been demonstrated (Miyata and Yasunaga, 1978; Pavesi, 2000; 2006; Jordan et al., 2000; Fujii et al. 2001). This extra evolutionary constraint exists due to the fact that a favorable or even neutral substitution in one reading frame could prove harmful for the other reading frame. Therefore, even a synonymous, favorable, or neutral nucleotide substitution in one reading frame might be discarded, as it could be deleterious in the other reading

frame. As a result, positive selection of overlapping genes in general is severely restricted (Miyata and Yasunaga, 1978; Mizokami et al., 1997; Pavesi, 2000; 2006; Jordan et al., 2000; Fujii et al. 2001). However, a recent study on HBV has proposed a new mechanism of evolution of overlapping genes termed “independent adaptive selection” (Zaaijer et al., 2007). In SARS-CoV, the orf9b gene overlaps completely with the nucleocapsid (N) gene (Fig. 10). Here, it was demonstrated that the overlapping region of the N gene is rather evolutionary conserved as compared to the orf9b gene. Orf9b features a higher evolutionary rate that is attained mainly via N3/9b2 substitutions (Shukla and Hilgenfeld, 2015). This mechanism of independent evolution is similar to the evolution described for the HBV surface protein gene, which completely overlaps with the polymerase gene (Zaaijer et al., 2007).

7 Conclusions

Viruses are extremely diverse organisms which evolve over a long period of time along and within their hosts. They make use of every imaginable genome type and replication strategy to survive (Watkiss, 2010). The pressure to adapt to new environmental conditions (hosts) and to replicate quickly and efficiently is very high in case of RNA viruses (Belshaw et al. 2007). The high mutation rate resulting from viral RNA polymerases gives RNA viruses the advantage to quickly adapt to environmental changes, but this also limits their genome size. This size limit results in a condensed genome with maximum information content. Hence, RNA viruses employ various strategies to maximize their coding potential. One of the mechanisms is the use of a secondary start codon in an alternating reading frame which results in the translation of different protein products. The findings of the study presented here contribute to characterizing sequence properties of newly acquired genes formed by overlapping reading frames. This will eventually lead to a better understanding of evolution.

Characterizing viral accessory proteins became even more of a challenge because of poor annotation; for example, in case of the SARS-CoV orf9b gene (Chapter 3). Due to the limitations of gene-finding algorithms, these overlapping coding regions are sometimes overlooked in the genome (Pavesi et al., 2013). However, new techniques are being developed based on codon usage, inherent disorder prediction, identification of conserved secondary and tertiary structural elements in the predicted genes (Rancurrel et al., 2009; Pavesi et al., 2013).

In this thesis, individual cases of overlapping viral genes were analysed using the techniques mentioned above. The codon usage pattern in one of the two overlapping viral proteins was found to be different from that of the rest of the genome. It was observed that a concurrent codon usage (with respect to the rest of the genome) occurred in the overprinted protein, that is the ancestral protein. So in conclusion, overprinting proteins with a discordant codon usage pattern have been more recently acquired in the course of viral evolution.

Inherent disorder prediction and structure comparison of the overlapping protein set showed that overlapping proteins tend to be disordered. In the region of overlap, the predicted disorder content is found to be higher when compared to the non-overlapping region. Inherently disordered proteins are believed to escape the evolutionary constraint imposed on their sequence by the overlap (Rancurrel et al., 2009).

To understand the molecular mechanism of translation of overlapping proteins, the RNA secondary structure around the initiation site of the alternative reading frame was probed. However, neither a consensus sequence, nor consensus secondary- or tertiary-RNA structural elements were identified among the overlapping genes. Therefore, it can be concluded that the RNA secondary structure surrounding the translation initiation site of the alternative reading frames are case-specific for an overlapping gene set.

It has been proposed that the choice of phase depends on the mutational load on the overlapping protein set (Normark et al., 1983; Johnson and Chrisholm, 2004). A recent study on overlapping gene sets of prokaryotes established a relationship between the frequency of mutation and the choice of optimal

phase (Lillo and Krakauer, 2007). A sufficient number of sequences was available in the overlapping gene set of SARS-CoV N and orf9b to make meaningful analyses and to find out the mutational model adopted by this protein set. A +1 frameshift at the 10^{th} nucleotide position of the N gene results in the translation of the orf9b protein. This leads to an interconnection between its second non-degenerate codon positions with the third, degenerate codon positions of the N protein. It was found that the recently acquired orf9b evolves independently of the overprinted nucleocapsid protein and that the relatively high mutability is indeed achieved by point mutations at the N3/9b2 codon positions (Chapter 6).

Bibliography

Alonso S, Izeta A, Sola I, Enjuanes L (2002): Transcription regulatory sequences and mRNA expression levels in the coronavirus transmissible gastroenteritis virus. J Virol. 76: 1293-1308.

Annan A, Baldwin HJ, Corman VM, Klose SM, Owusu M, Nkrumah EE, Badu EK, Anti P, Agbenyega O, Meyer B, Oppong S, Sarkodie YA, Kalko EK, Lina PH, Godlevska EV, Reusken C, Seebens A, Gloza-Rausch F, Vallo P, Tschapka M, Drosten C, Drexler JF (2013): Human betacoronavirus 2c EMC/2012-related viruses in bats, Ghana and Europe. Emerg. Infect. Dis. 19: 456-459.

Aota S, Ikemura T (1986): Diversity in G + C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res. 14: 6345-6355.

Barrell BG, Air GM, Hutchison CA (1976): Overlapping genes in bacteriophage phiX174. Nature 264: 34-41.

Beier H (1984): UAG readthrough during TMV RNA translation: isolation and sequence of two tRNAs with suppressor activity from tobacco plants. EMBO J. 3**:** 351-356.

Belshaw R, Pybus OG, Rambaut A (2007): The evolution of genome compression and genomic novelty in RNA viruses. Genome Res. 10: 1496-1504.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2012): GenBank. Nucleic Acids Res. 41: 36-42.

Brierely I, Dos Ramos FJ (2006): Programmed ribosomal frameshifting in HIV-1 and the SARS-CoV. Virus Res. 119**:** 29-42.

Brierley I, Digard P, Inglis S (1989): Characterization of an efficient ribosomal frameshifting signal: Requirement for an RNA pseudoknot. Cell 57: 537-547.

Brown EA, Zhang H, Ping L, Lemon SM (1992): Secondary structure of the 5′ nontranslated regions of hepatitis C virus and pestivirus genomic RNAs. Nucleic Acids Res. 20: 5041–5045.

Casais R, Davies M, Cavanagh D, Britton P (2005): Gene 5 of the avian coronavirus infectious bronchitis virus is not essential for replication. J. Virol. 79: 8065-8078.

Castaño A, Ruiz L, Hernández C (2009): Insights into the translational regulation of biologically active open reading frames of Pelargonium line pattern virus. Virology 38: 417–426.

Chan WS, Wu C, Chow SC, Cheung T, To KF, Leung WK, Chan PK, Lee KC, Ng HK, Au DM, Lo AW (2005): Coronaviral hypothetical and structural proteins were found in the intestinal surface enterocytes and pneumocytes of severe acute respiratory syndrome (SARS). Mod. Pathol. 18: 1432-1439.

Chan CM, Tsoi H, Chan WM, Zhai S, Wong CO, Yao X, Chan WY, Tsui SK, Chan HY (2009): The ion channel activity of the SARS-coronavirus 3a protein is linked to its pro-apoptotic function. Int. J. Biochem. Cell Biol. 41: 2232-2239.

Chang C, Hou MH, Chang CF, Hsiao CD, Huang TH (2014): The SARS coronavirus nucleocapsid protein- forms and functions. Antiviral Res. 103: 39-50.

Chen P, Gan Y, Han N, Fang W, Li J, Zhao F, Hu K, Rayner S (2013): Computational evolutionary analysis of the overlapped surface (S) and polymerase (P) region in hepatitis B virus indicates the spacer domain in P is crucial for survival. PLoS One 8:e60098.

Cheng VCC, Chan JFW, To KKW, Yuen KY (2013): Clinical management and infection control of SARS. Antiviral Res. 100: 407-419.

Chirico N, Vianelli A, Belshaw R (2010): Why genes overlap in viruses? Proc. Biol. Sci. 277: 3809-3817.

Cotten M, Lam TT, Watson SJ, Palser AL, Petrova V, Grant P, Pybus OG,

Rambaut A, Guan Y, Pillay D, Kellam P, Nastouli E (2013): Full-genome deep sequencing and phylogenetic analysis of novel human betacoronavirus. Emerg. Infect. Dis. 5: 736-742.

Crick FH (1968): The origin of the genetic code. J. Mol. Biol. 38: 367–379.

de Groot RJ, Gorbalenya AE (2010): ICTV proposal to include a new genus and three new species in the subfamily Coronavirinae. (online available at http://ictvonline.org/proposals/2008.085-122V.v4.Coronaviridae.pdf,last aces--sed on 26.02.2015).

de Groot RJ, Baker SC, Baric R, Enjuanes L, Gorbalenya AE, Holmes KV, Perlman S, Poon L, Rottier PJM, Talbot PJ, Woo PCY, Ziebuhr J (2011): *Coronaviridae*, p 806 – 828. *In* King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (ed), Virus taxonomy: Ninth report of the International Committee on Taxonomy of Viruses, International Union of Microbiological Societies, Virology Division. Elsevier Academic Press, London, United Kingdom.

de Groot RJ, Baker SC, Baric RS, Brown CS, Drosten C, Enjuanes L, Fouchier RA, Galiano M, Gorbalenya AE, Memish ZA, Perlman S, Poon LL, Snijder EJ, Stephens GM, Woo PC, Zaki AM, Zambon M, Ziebuhr J (2013): Middle East Respiratory Syndrome coronavirus (MERS-CoV); Announcement of the Coronavirus Study Group. J. Virol. 87: 7790-7792.

Dedeurwaerder A, Desmarets LM, Olyslaegers DA, Vermeulen BL, Dewerchin HL, Nauwynck HJ (2013): The role of accessory proteins in the replication of feline infectious peritonitis virus in peripheral blood monocytes. Vet Microbiol. 162: 447-455.

Dedeurwaerder A, Olyslaegers DA, Desmarets LM, Roukaerts ID, Theuns S, Nauwynck HJ (2014): The ORF7-encoded accessory protein 7a of feline infectious peritonitis virus as a counteragent against interferon-alpha induced antiviral response. J. Gen. Virol. 95: 393-402.

Dijkman R, Jebbink MF, Wilbrink B, Pyrc K, Zaaijer HL, Minor PD, Franklin S, Berkhout B, Thiel V, van der Hoek L (2006): Human

coronavirus 229E encodes a single ORF4 protein between the spike and the envelope genes. Virol J. 3: 106.

Dinman J (2012): Mechanisms and implications of programmed translational frameshifting. Wiley Interdiscip. Rev. RNA 3: 661-673.

Dreher TW, Miller WA (2006): Translational control in positive strand RNA plant viruses. Virology 344: 185-197.

Drexler JF, Corman VM, Drosten C (2014): Ecology, evolution and classification of bat coronaviruses in the aftermath of SARS. Antiviral Res. 101: 45-56.

Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z (2001): Intrinsically disordered protein. J. Mol. Graph. Model. 19: 26-59.

Dunker AK, Silman I, Uversky VN, Sussman JL (2008): Function and structure of inherently disordered proteins. Curr. Opin. Struct. Biol. 18: 756-764.

Dyson HJ, Wright PE (2005): Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. 6: 197-208.

Farsani SMJ, Dijkman R, Jebbink MF, Goossens H, Ieven M, Deijs M, Molenkamp R, van der Hoek L (2012): The first complete genome sequences of clinical isolates of human coronavirus 229E. Virus Genes 45: 433-439.

Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA (eds) (2005): Virus Taxonomy, Eighth Report of the International Committee on Taxonomy of Viruses. Elsevier/Academic Press, London. (online access at http://ictvonline.org, last accessed on 26.02.2015).

Firth AE, Brierley I (2012): Non-canonical translation in RNA viruses. J. Gen. Virol. 93: 1385-1409.

Fischer F, Peng D, Hingley ST, Weiss SR, Masters PS (1997): The internal open reading frame within the nucleocapsid gene of mouse hepatitis virus encodes a structural protein that is not essential for viral replication. J. Virol. 71: 996-1003.

Forsdyke DR (2012): Grantham's genome hypotheses. Available from webpage http://post.queensu.ca/~forsdyke/granth01.htm, created in 2000 and last edited on 21st September 2012 (last accessed on 18.02.2015).

Fredslund J (2006): PHY.FI: fast and easy online creation and manipulation of phylogeny color figures. BMC Bioinformatics 7: 315.

Freundt EC, Yu L, Park E, Lenardo MJ, Xu XN (2009): Molecular determinants for subcellular localization of the severe acute respiratory syndrome coronavirus open reading frame 3b protein. J. Virol. 83: 6631-6640.

Freundt EC, Yu L, Goldsmith CS, Welsh S, Cheng A, Yount B, Liu W, Frieman MB, Buchholz UJ, Screaton GR, Lippincott-Schwartz J, Zaki SR, Xu XN, Baric RS, Subbarao K, Lenardo MJ (2010): The open reading frame 3a protein of severe acute respiratory syndrome-associated coronavirus promotes membrane rearrangement and cell death. J. Virol. 84: 1097-1109.

Fujii Y, Kiyotani K, Yoshida T, Sakaguchi T (2001): Conserved and non-conserved regions in the Sendai virus genome: evolution of a gene possessing overlapping reading frames. Virus Genes 22**:** 47-52.

Fukuda Y, Washio T, Tomita M (1999): Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. Nucleic Acids Res. 27: 1847‑1853.

Fukuda Y, Nakayama Y, Tomita M (2003): On dynamics of overlapping genes in bacterial genomes. Gene 323: 181‑187.

Ge XY, Li JL, Yang XL, Chmura AA, Zhu G, Epstein JH, Mazet JK, Hu B, Zhang W, Peng C, Zhang YJ, Luo CM, Tan B, Wang N, Zhu Y, Crameri G, Zhang SY, Wang LF, Daszak P, Shi ZL (2013): Isolation and

characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature 503: 535-538.

Grantham R, Gautier C, Gouy M, Mercier R, Pavé A (1980): Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8: 49-62.

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981): Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9: 43-74.

Gu J, Gribskov M, Bourne PE (2006): Wiggle-predicting functionally flexible regions from primary sequence. PLoS Comput. Biol. 2: e90.

Gu J, Hilser V (2009): The significance and impacts of protein disorder and conformational variants. In Structural Bioinformatics, Second Ed. (edited by Gu J, Bourne PE), ISBN: 978-0-0470-18105-8, Wiley & Blackwell, pp. 939-962.

Guo JP, Petric M, Campbell W, McGeer PL (2004): SARS corona virus peptides recognized by antibodies in the sera of convalescent cases. Virology 324: 251–256.

Haagmans BL, Al Dhahiry SH, Reusken CB, Raj VS, Galiano M, Myers R, Godeke GJ, Jonges M, Farag E, Diab A, Ghobashy H, Alhajri F, Al-Thani M, Al-Marri SA, Al Romaihi HE, Al Khal A, Bermingham A, Osterhaus AD, AlHajri MM, Koopmans MP (2014): Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation. Lancet Infect. Dis. 14: 140–145.

Haijema BJ, Volders H, Rottier PJ (2004): Live, attenuated coronavirus vaccines through the directed deletion of group-specific genes provide protection against feline infectious peritonitis. J. Virol. 78: 3863-3871.

Hall TA (1999): BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41**:** 95–98.

Hilgenfeld R, Peiris JSM (2013): From SARS to MERS: 10 years of research on highly pathogenic human coronaviruses. Antiviral Res. 100: 286-295.

Hodgson T, Britton P, Cavanagh D (2006): Neither the RNA nor the proteins of open reading frames 3a and 3b of the coronavirus infectious bronchitis virus are essential for replication. J. Virol. 80: 296-305.

Honigman A (1991): *cis* acting RNA sequences control the gag-pol translation readthrough in murine leukemia virus. Virology 183: 313-319.

Hurst LD (2002): The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet 8: 486.

Ikemura T (1985): Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 2: 13-34.

Ithete NL, Stoffberg S, Corman VM, Cottontail VM, Richards LR, Schoeman MC, Drosten C, Drexler JF, Preiser W (2013): Close relative of human Middle East respiratory syndrome coronavirus in bat, South Africa. Emerg. Infect. Dis. 19: 1697–1699

Ito N, Mossel EC, Narayanan K, Popov VL, Huang C, Inoue T, Peters CJ, Makino S (2005): Severe acute respiratory syndrome coronavirus 3a protein is a viral structural protein. J. Virol. 79: 3182-3186.

Jacks T, Madhani HD, Masiarz FR, Varmus HE (1988): Signals for ribosomal frameshifting in Rous Sarcoma virus gag-pol region. Cell 55**:** 447-458.

Jackson RJ, Howell MT, Kaminski A (1990): The novel mechanism of initiation of picornavirus RNA translation. Trends Biochem. Sci. 15: 477-483.

Jackson RJ, Kaminski A (1995): Internal initiation of translation in eukaryotes: the picornavirus paradigm and beyond. RNA 1: 985-1000.

Jackson RJ (2000): A Comparative View of Initiation Site Selection Mechanisms. In Translational control of gene expression (edited by: Nahum Sonenberg, John WB Hershey and Michael B Mathews), ISBN: 0-87969-618-

4, Cold Spring Harbor Laboratory Press, USA, pp: 127-183.

Jayalakshmi MK, Kalyanaraman N, Pitchappan R (2013): Hepatitis B Virus genetic diversity: Disease Pathogenesis. In Viral Replication (edited by: German Rosas-Acosta), ISBN: 978-953-51-1055-2, InTech, available online at: http://www.intechopen.com/books/viral-replication/hepatitis-b-virus-genetic-diversity-disease-pathogenesis (last accessed on 02.03.2015).

Johnson ZI, Chisholm SW (2004): Properties of overlapping genes are conserved across microbial genomes. Genome Res. 14: 2268-2272.

Jones DT, Ward JJ (2003): Prediction of disordered regions in proteins from position specific score matrices. Proteins 53: 573-578.

Jordan KI, Sutter BA, McClure MA (2000): Molecular evolution of the paramyxoviridae and the rhabdoviridae multiple-protein-encoding P gene. Mol. Biol. Evol. 17**:** 75-86.

Karlin D, Ferron F, Canard B, Longhi S (2003): Structural disorder and modular organization in Paramyxovirinae N and P. J. Gen. Virol. 84: 3239-3252.

Karst SM, Wobus CE, Lay M, Davidson J, Virgin HW (2003): STAT1-dependent innate immunity to a Norwalk-like virus. Science 299: 1575-1578.

Keese PK, Gibbs A (1992): Origins of genes: "big bang" or continuous creation? Proc. Natl. Acad. Sci. USA 89: 9489–9493.

Khan S, Fielding BC, Tan TH, Chou C-F, Shen S, Lim SG, Hong W, Tan YJ (2006): Over-expression of severe acute respiratory syndrome coronavirus 3b protein induces both apoptosis and necrosis in Vero E6 cells. Virus Res. 122: 20-27.

Kieft JS (2008): Viral IRES RNA structures and ribosome interactions. Trends Biochem. Sci. 33: 274-283.

Koonin EV, Novozhilov AS (2009): Origin and evolution of the genetic code: the universal enigma. IUBMB Life. 61: 99-111

Kopecky-Bromberg SA, Martinez-Sobrido L, Frieman M, Baric RA, Palese P (2007): Severe acute respiratory syndrome coronavirus open reading frame (ORF) 3b, ORF 6, and nucleocapsid proteins function as interferon antagonists. J. Virol. 81: 548-557.

Kozak M (1983): Translation of insulin-related polypeptides from messenger RNAs with tandemly reiterated copies of the ribosome binding site. Cell 34: 971–978.

Kozak M (1986): Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 44: 283–292.

Kozak M (1989): The scanning model for translation: an update. J. Cell Biol. 108: 229-241.

Kozak M (1995): Adherence to the first-AUG rule when a second AUG codon follows closely upon the first. Proc. Natl. Acad. Sci. USA 92: 2662-2666.

Kozak M (1997): Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. EMBO J. 16: 2482-2492.

Kozak M (1999): Initiation of translation in prokaryotes and eukaryotes. Gene 234**:** 187-208.

Kozak M (2002): Pushing the limits of the scanning mechanism for initiation of translation. Gene 299: 1–34.

Kozak M (2005): Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene 361: 13-37.

Krakauer DC (2000): Stability and evolution of overlapping genes. Evolution 54**:** 731-739.

Krakauer DC, Plotkin JB (2002): Redundancy, antiredundancy, and the robustness of genomes. Proc. Natl. Acad. Sci. USA 99: 1405-1409.

Langereis MA, Zeng Q, Heesters B, Huizinga EG, de Groot RJ (2012): The murine coronavirus hemagglutinin-esterase receptor-binding site: a major shift in ligand specificity through modest changes in architecture. PLoS Pathog. 8: e1002492.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007): Clustal W and Clustal X version 2.0. Bioinformatics 23**:** 2947-2948.

Lassmann T, Sonnhammer EL (2005): Kalign--an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics 6: 298.

Lau SK, Poon RW, Wong BH, Wang M, Huang Y, Xu H, Guo R, Li KS, Gao K, Chan KH, Zheng BJ, Woo PC, Yuen KY (2010): Coexistence of different genotypes in the same bat and serological characterization of *Rousettus* bat coronavirus HKU9 belonging to a novel Betacoronavirus subgroup. J. Virol. 84: 11385-11394.

Law PT, Wong CH, Au TC, Chuck CP, Kong SK, Chan PK, To KF, Lo AW, Chan JY, Suen YK, Chan HY, Fung KP, Waye MM, Sung JJ, Lo YM, Tsui SK (2005): The 3a protein of severe acute respiratory syndrome-associated coronavirus induces apoptosis in Vero E6 cells. J. Gen. Virol. 86: 1921-1930.

Levin DB, Whittome B (2000): Codon usage in nucleopolyhedroviruses. J. Gen. Virol. 81: 2313-2325.

Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, Wang H, Crameri G, Hu Z, Zhang H, Zhang J, McEachern J, Field H, Daszak P, Eaton BT, Zhang S, Wang LF (2005): Bats are natural reservoirs of SARS-like coronaviruses. Science 310: 676-679.

Librado P, Rozas J (2009): DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25**:** 1451-1452.

Lillo F, Krakauer DC (2007): A statistical analysis of the three-fold evolution of genomic compression through frame overlaps in prokaryotes. Biol. Dir. 2:

22.

Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB (2003a): Protein disorder prediction: implications for structural proteomics. Structure 11: 1453-1459.

Linding R, Russell RB, Neduva V, Gibson TJ (2003b): GlobPlot: exploring protein sequences for globularity and disorder. Nucleic Acids Res. 31: 3701-3708.

Liu DX, Inglis SC (1992): Internal entry of ribosomes on a tricistronic mRNA encoded by infectious bronchitis virus. J. Virol. 66: 6143-6154.

Liu DX, Fung TS, Chong KK, Shukla A, Hilgenfeld R (2014): Accessory proteins of SARS-CoV and other coronaviruses. Antiviral Res. 109: 97-109.

Lorenz R, Bernhart SH, zu Siederdissen CH, Tafer H, Flamm C, Stadler PF, Hofacker IL (2011): ViennaRNA Package 2.0. Algorithms Mol. Biol. 6: 26.

Lu W, Zheng BJ, Xu K, Schwarz W, Du L, Wong CK, Chen J, Duan S, Deubel V, Sun B (2006): Severe acute respiratory syndrome-associated coronavirus 3a protein forms an ion channel and modulates virus release. Proc. Natl. Acad. Sci. USA 103: 12540-12545.

Luytjes W, Bredenbeek PJ, Noten AF, Horzinek MC, Spaan WJ (1988): Sequence of mouse hepatitis virus A59 mRNA 2: indications for RNA recombination between coronaviruses and influenza C virus. Virology 166: 415-422.

Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YS, Khattra J, Asano JK, Barber SA, Chan SY, Cloutier A, Coughlin SM, Freeman D, Girn N, Griffith OL, Leach SR, Mayo M, McDonald H, Montgomery SB, Pandoh PK, Petrescu AS, Robertson AG, Schein JE, Siddiqui A, Smailus DE, Stott JM, Yang GS, Plummer F, Andonov A, Artsob H, Bastien N, Bernard K, Booth TF, Bowness D, Czub M, Drebot M, Fernando L, Flick R, Garbutt M, Gray M, Grolla A, Jones S, Feldmann H, Meyers A, Kabani A, Li Y, Normand S, Stroher U, Tipples GA, Tyler S,

Vogrig R, Ward D, Watson B, Brunham RC, Krajden M, Petric M, Skowronski DM, Upton C, Roper RL (2003): The genome sequence of the SARS-associated coronavirus. Science 300: 1399-1404.

Marzi S, Myasnikov AG, Serganov A, Ehresmann C, Romby P, Yusupov M, Klaholz BP (2007): Structured mRNAs regulate translation initiation by binding to the platform of the ribosome. Cell 130: 1019-1031.

Masters PS (2006): The molecular biology of coronaviruses. Adv. Virus Res. 66: 193-292.

Matsuda D, Dreher TW (2006): Close spacing of AUG initiation codons confers dicistronic character on a eukaryotic mRNA. RNA 12**:** 1338-1349.

Matthews KL, Coleman CM, van der Meer Y, Snijder EJ, Frieman MB (2014): The ORF4b-encoded accessory proteins of Middle East respiratory syndrome coronavirus and two related bat coronaviruses localize to the nucleus and inhibit innate immune signalling. J. Gen. Virol. 95: 874-882.

McBride R, Fielding BC (2012): The role of severe acute respiratory syndrome (SARS)-coronavirus accessory proteins in virus pathogenesis. Viruses 4: 2902-2923.

McFadden N, Bailey D, Carrara G, Benson A, Chaudhry Y, Shortland A, Heeney J, Yarovinsky F, Simmonds P, Macdonald A, Goodfellow I (2011): Norovirus regulation of the innate immune response and apoptosis occurs via the product of the alternative open reading frame 4. PLoS Pathog. 7: e1002413.

McInerney JO (1998): Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc. Natl. Acad. Sci. USA. 95:10698-10703.

Meier C, Aricescu AR, Stuart DI, Grimes J, Gilbert RJC, Aplin RT, Assenberg R (2006): The crystal structure of ORF 9b, a lipid binding protein from the SARS Coronavirus. Structure 14: 1157-1165.

Meyer B, Müller MA, Corman VM, Reusken CB, Ritz D, Godeke GJ, Lattwein E, Kallies S, Siemens A, van Beek J, Drexler JF, Muth D, Bosch

BJ, Wernery U, Koopmans MP, Wernery R, Drosten C (2013): Antibodies against MERS coronavirus in dromedary camels, United Arab Emirates, 2003 and 2013. Emerg. Infect. Dis. 20: 552–559.

Miyata T, Yasunaga T (1978): Evolution of overlapping genes. Nature 272: 532-535.

Mizokami M, Orito E, Ohba K, Ikeo K, Lau JY, Gojobori T (1997): Constrained evolution with respect to gene overlap of hepatitis B virus. J. Mol. Evol. 44(Suppl 1)**:** S83-S90.

Moshynskyy I, Viswanathan S, Vasilenko N, Lobanov V, Petric M, Babiuk LA, Zakhartchouk AN (2007): Intracellular localization of the SARS coronavirus protein 9b: evidence of active export from the nucleus. Virus Res. 127: 116-121.

Narayanan K, Huang C, Makino S (2008): SARS coronavirus accessory proteins. Virus Res. 133: 113-121.

Niemeyer D, Zillinger T, Muth D, Zielecki F, Horvath G, Suliman T, Barchet W, Weber F, Drosten C, Müller MA (2013): Middle East respiratory syndrome coronavirus accessory protein 4a is a type I interferon antagonist. J. Virol. 87: 12489-12495.

Normark S, Bergström S, Edlund T, Grundström T, Jaurin B, Lindberg FP, Olsson O (1983): Overlapping genes. Annu. Rev. Genet 17: 499-525.

Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK (2003): Predicting intrinsic disorder from amino acid sequence. Proteins 53**:** 566-572.

Orlova M (2003): Reverse transcriptase of Moloney Murine Leukemia Virus binds to eukaryotic release factor 1 to modulate suppression of translational termination. Cell 115**:** 319-331.

Paul PS, Vaughn EM, Halbur PG (1997): Pathogenicity and sequence analysis studies suggest potential role of gene 3 in virulence of swine enteric and respiratory coronaviruses. Adv. Exp. Med. Biol. 412: 317-321.

Pavesi A, De Iaco B, Granero MI, Porati A (1997): On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. J. Mol. Evol. 44: 625-631.

Pavesi A (2000): Detection of signature sequences in overlapping genes and prediction of a novel overlapping gene in hepatitis G virus. J. Mol. Evol. 50: 284-295.

Pavesi A (2006): Origin and evolution of overlapping genes in the family *Microviridae*. J. Gen. Virol. 87**:** 1013-1017.

Pavesi A, Magiorkinis G, Karlin DG (2013): Viral proteins originated *de novo* by overprinting can be identified by codon usage: application to the "gene nursery" of deltaretroviruses. PLoS Comput. Biol. 9: e1003162.

Peden JF (1999): Analysis of codon usage. PhD Thesis, University of Nottingham, UK, available online at http://codonw.sourceforge.net/John PedenThesisPressOpt_water.pdf (Last accessed on 15.04.2015).

Plant EP (2012): Ribosomal frameshift signals in viral genomes. In Viral Genomes - Molecular structure, diversity, gene expression mechanisms and host-virus interactions (edited by M. Garcia), ISBN: 978-953-51-0098-0, InTech, available online at: http://www.intechopen.com/books/viral-genomes-molecular-structure-diversity-gene-expression-mechanisms-and-host-virus-interactions/frameshift-signals-in-viral-genomes (Last accessed on 01.03.2015).

Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, Blazewicz J, Adamiak RW (2012): Automated 3D structure composition for large RNAs. Nucleic Acids Res. 40: e112.

Qiu M, **Shi Y**, **Guo Z**, **Chen Z**, **He R**, **Chen R**, **Zhou D**, **Dai E**, **Wang X**, **Si B**, **Song Y**, **Li J**, **Yang L**, **Wang J**, **Wang H**, **Pang X**, **Zhai J**, **Du Z**, **Liu Y**,

Zhang Y, **Li L**, **Wang J**, **Sun B**, **Yang R** (2005): Antibody responses to individual proteins of SARS coronavirus and their neutralization activities. Microb. Infect. 7: 882-889.

Rancurrel C, Khosravi M, Dunker AK, Romero P, Karlin D (2009): Overlapping genes produce proteins with unusual sequence properties and offer insight into *de novo* protein creation. J. Virol. 83: 10719-10736.

Reusken CB, Haagmans BL, Müller MA, Gutierrez C, Godeke GJ, Meyer B, Muth D, Raj VS, Smits-De Vries L, Corman VM, Drexler JF, Smits SL, El Tahir YE, De Sousa R, van Beek J, Nowotny N, van Maanen K, Hidalgo-Hermoso E, Bosch BJ, Rottier P, Osterhaus A, Gortázar-Schmidt C, Drosten C, Koopmans MP (2013): Middle East respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: a comparative serological study. Lancet Infect. Dis. 13: 859–866

Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, Tatusov RL, Koonin EV (2002): Purifying and directional selection in overlapping prokaryotic genes. Trends Genet 18: 228-232.

Romero P, Obradovic Z, Kissinger C, Villafranca JE, Dunker AK (1997): Identifying disordered regions in proteins from amino acid sequences. Proc. IEEE. Int. Conf. Neural Networks 1 : 90-95.

Romero P, Obradovic Z, Li XH, Garner EC, Brown CJ, Dunker AK (2001): Sequence complexity of disordered protein. Proteins 42: 38-48.

Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, Peñaranda S, Bankamp B, Maher K, Chen MH, Tong S, Tamin A, Lowe L, Frace M, DeRisi JL, Chen Q, Wang D, Erdman DD, Peret TC, Burns C, Ksiazek TG, Rollin PE, Sanchez A, Liffick S, Holloway B, Limor J, McCaustland K, Olsen-Rasmussen M, Fouchier R, Günther S, Osterhaus AD, Drosten C, Pallansch MA, Anderson LJ, Bellini WJ (2003): Characterization of a novel coronavirus associated with Severe Acute Respiratory Syndrome. Science 300: 1394-1399.

Ryabova LA, Pooggin MM, Hohn T (2006): Translation reinitiation and leaky scanning in plant viruses. Virus Res. 119**:** 52-62.

Saikatendu KS, Joseph JS, Subramanian V, Neuman BW, Buchmeier MJ, Stevens RC, Kuhn P (2007): Ribonucleocapsid formation of severe acute respiratory syndrome coronavirus through molecular action of the N-terminal domain of N protein. J. Virol. 81: 3913-3921.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (1977): Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265: 687-695.

Sawicki SG, Sawicki DL, Siddell SG (2007): A contemporary view of coronavirus transcription. J. Virol. 81: 20–29.

Sharma K, Åkerström S, Sharma AK, Chow VT, Teow S, Abrenica B, Booth SA, Booth TF, Mirazimi A, Lal SK (2011): SARS-CoV 9b protein diffuses into nucleus, undergoes active Crm1 mediated nucleocytoplasmic export and triggers apoptosis when retained in the nucleus. PLoS ONE 6: e19436.

Sharp PM, Li WH (1987): The codon adaptation index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15: 1281-1295.

Sharp PM, Emery LR, Zeng K (2010): Forces that influence the evolution of codon bias. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 365: 1203-1212.

Shen S, Lin PS, Chao YC, Zhang A, Yang X, Lim SG, Hong W, Tan YJ (2005): The severe acute respiratory syndrome coronavirus 3a is a novel structural protein. Biochem. Biophys. Res. Commun. 330: 286-292.

Shukla A, Hilgenfeld R (2015): Acquisition of new protein domains by coronaviruses: analysis of overlapping genes coding for proteins N and 9b in SARS coronavirus. Virus Genes 50: 29-38.

Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A,

Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK (2007): DisProt: the database of disordered proteins. Nucleic Acids Res. 35: 786-793.

Siu KL, Yeung ML, Kok KH, Yuen KS, Kew C, Lui PY, Chan CP, Tse H, Woo PC, Yuen KY, Jin DY (2014): Middle east respiratory syndrome coronavirus 4a protein is a double-stranded RNA-binding protein that suppresses PACT-induced activation of RIG-I and MDA5 in the innate antiviral response. J. Virol. 88: 4866-4876.

Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, Gorbalenya AE (2003): Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. J. Mol. Biol. 331: 991-1004.

Stothard P (2000): The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 28:1102-1104.

Tan YJ, Teng E, Shen S, Tan TH, Goh PY, Fielding BC, Ooi EE, Tan HC, Lim SG, Hong W (2004): A novel severe acute respiratory syndrome coronavirus protein, U274, is transported to the cell surface and undergoes endocytosis. J. Virol. 78: 6723-6734.

Tan TH, Barkham T, Fielding BC, Chou CF, Shen S, et al Lim SG, Hong W, Tan YJ (2005): Genetic lesions within the 3a gene of SARS-CoV. Virol. J. 2: 51.

Tan YJ, Lim SG, Hong W (2006): Understanding the accessory viral proteins unique to the severe acute respiratory syndrome (SARS) coronavirus. Antiviral Res. 72: 78-88.

Thiel V, Siddell SG (1994): Internal ribosome entry in the coding region of murine hepatitis virus mRNA 5. J. Gen. Virol. 75: 3041-3046.

Tung FY, Abraham S, Sethna M, Hung S-L, Sethna P, Hogue BG, Brian DA (1992): The 9-kDa hydrophobic protein encoded at the 3′ end of the

porcine transmissible gastroenteritis coronavirus genome is membrane-associated. Virology 186: 676-683.

van Boheemen S, de Graaf M, Lauber C, Bestebroer TM, Raj VS, Zaki AM, Osterhaus AD, Haagmans BL, Gorbalenya AE, Snijder EJ, Fouchier RA (2012): Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. Mbio 3: e00473-00412.

Wang X, Wong SM, Liu DX (2006): Identification of hepta- and octo-uridine stretches as sole signals for programmed +1 and -1 ribosomal frameshifting during translation of SARS-CoV ORF 3a variants. Nucleic Acids Res. 34: 1250-1260.

Watkiss E (2010): RNA viruses: Strategies to maximize coding potential. In online article available at: http://homepage.usask.ca/~vim458/advirol/ 2010/watkiss/viral_strategies.pdf (last accessed on 18.02.2015).

Woo PC, Wang M, Lau SK, Xu H, Poon RW, Guo R, Wong BH, Gao K, Tsoi HW, Huang Y, Li KS, Lam CS, Chan KH, Zheng BJ, Yuen KY (2007): Comparative analysis of twelve genomes of three novel group 2c and group 2d coronaviruses reveals unique group and subgroup features. J. Virol. 81: 1574-1585.

Woo PC, Lau SK, Lam CS, Lai KK, Huang Y, Lee P, Luk GS, Dyrting KC, Chan KH, Yuen KY (2009): Comparative analysis of complete genome sequences of three avian coronaviruses reveals a novel group 3c coronavirus. J Virol. 83: 908-917.

Woo PC, Huang Y, Lau SK, Yuen KY (2010): Coronavirus genomics and bioinformatics analysis. Viruses 2: 1804-1820.

Woo PC, Lau SK, Lam CS, Lau CC, Tsang AK, Lau JH, Bai R, Teng JL, Tsang CC, Wang M, Zheng BJ, Chan KH, Yuen KY (2012): Discovery of seven novel Mammalian and avian coronaviruses in the genus *Deltacoronavirus* supports bat coronaviruses as the gene source of *Alphacoronavirus* and *Betacoronavirus* and avian coronaviruses as the gene

source of *Gammacoronavirus* and *Deltacoronavirus*. J Virol. 86: 3995-4008.

Woo PC, Lau SK, Lam CS, Tsang AK, Hui SW, Fan RY, Martelli P, Yuen KY (2014): Discovery of a novel bottlenose dolphin coronavirus reveals a distinct species of marine mammal coronavirus in *Gammacoronavirus*. J Virol. 88: 1318-1331.

Wootton JC, Federhen S (1993): Statistics of local complexity in amino-acid-sequences and sequence databases. Comp. Chem. 17: 149-163.

Wootton JC, Federhen S (1996): Analysis of compositionally biased regions in sequence databases. Comp. Methods Macromol. Sequence Analys. 266: 554-571.

World Health Organization, Global Alert and Response (GAR). Middle East respiratory syndrome coronavirus (MERS-CoV) — summary updates, http://www.who.int/csr/don/9-april-2015-mers-saudi-arabia/en/ (last accessed on 15.04.2015).

Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z (2007): Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J. Proteome Res. 6**:**1882–1898.

Xu K, Zheng BJ, Zeng R, Lu W, Lin YP, Xue L, Li L, Yang LL, Xu C, Dai J, Wang F, Li Q, Dong QX, Yang RF, Wu JR, Sun B (2009): Severe acute respiratory syndrome coronavirus accessory protein 9b is a virion-associated protein. Virology 388: 279-285.

Yang ZR, Thomson R, McNeil P, Esnouf RM (2005): RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21: 3369-3376.

Yang Y, Zhang L, Geng H, Deng Y, Huang B, Guo Y, Zhao Z, Tan W (2013): The structural and accessory proteins M, ORF 4a, ORF 4b, and ORF 5 of Middle East respiratory syndrome coronavirus (MERS-CoV) are potent interferon antagonists. Protein Cell 4: 951-961.

Yang Y, Du L, Liu C, Wang L, Ma C, Tang J, Baric RS, Jiang S, Li F (2014): Receptor usage and cell entry of bat coronavirus HKU4 provide insight into bat-to-human transmission of MERS coronavirus. Proc. Natl. Acad. Sci. USA. 111: 12516–12521

Yu CJ, Chen YC, Hsiao CH, Kuo TC, Chang SC, Lu CY, Wei WC, Lee CH, Huang LM, Chang MF, Ho HN, Lee FJ (2004): Identification of a novel protein 3a from severe acute respiratory syndrome coronavirus. FEBS Lett. 565: 111-116.

Yu, I.M., Oldham, M.L., Zhang, J., Chen, J (2006): Crystal structure of the severe acute respiratory syndrome (SARS) coronavirus nucleocapsid protein dimerization domain reveals evolutionary linkage between corona- and arteriviridae. J.Biol.Chem. 281: 17134-17139.

Yuan X, Yao Z, Shan Y, Chen B, Yang Z, Wu J, Zhao Z, Chen J, Cong Y (2005): Nucleolar localization of non-structural protein 3b, a protein specifically encoded by the severe acute respiratory syndrome coronavirus. Virus Res. 114: 70-79.

Yuan X, Shan Y, Yao Z, Li J, Zhao Z, Chen J, Cong Y (2006): Mitochondrial location of severe acute respiratory syndrome coronavirus 3b protein. Mol. Cells 21: 186-191.

Zaaijer HL, van Hemert FJ, Koppelman MH, Lukashov VV (2007): Independent evolution of overlapping polymerase and surface protein genes of hepatitis B virus. J. Gen. Virol. 88: 2137–2143.

Zajakina A, Kozlovska T, Bruvere R, Aleksejeva J, Pumpens P, Garoff H (2004): Translation of hepatitis B virus (HBV) surface proteins from the HBV pregenome and precore RNAs in Semliki Forest virus-driven expression. J. Gen. Virol. 85: 3343-3351.

Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM (2012): Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N. Engl. J. Med. 367: 1814-1820.

Zeng R, Yang RF, Shi MD, Jiang MR, Xie YH, Ruan HQ, Jiang XS, Shi L, Zhou H, Zhang L, Wu XD, Lin Y, Ji YY, Xiong L, Jin Y, Dai EH, Wang XY, Si BY, Wang J, Wang HX, Wang CE, Gan YH, Li YC, Cao JT, Zuo JP, Shan SF, Xie E, Chen SH, Jiang ZQ, Zhang X, Wang Y, Pei G, Sun B, Wu JR (2004): Characterization of the 3a protein of SARS-associated coronavirus in infected vero E6 cells and SARS patients. J. Mol. Bio. 341: 271-279.

Zeng Q, Langereis MA, van Vliet ALW, Huizinga EG, de Groot RJ (2008): Structure of coronavirus hemagglutinin-esterase offers insight in corona and influenza virus evolution. Proc. Natl. Acad. Sci. USA 105: 9065–9069.

Zhang X, Kousoulas KG, Storz J (1992): The hemagglutinin/esterase gene of human coronavirus strain OC43: phylogenetic relationships to bovine and murine coronaviruses and influenza C virus. Virology 186: 318–323.

Zhang Y, Stec B, Godzik A (2007): Between order and disorder in protein structures: analysis of ‘‘dual personality’’ fragments in proteins. Structure 15: 1141-1147.

Zhong X, Guo Z, Yang H, Peng L, Xie Y, Wong TY, Lai ST, Guo Z (2006): Amino terminus of the SARS coronavirus protein 3a elicits strong, potentially protective humoral responses in infected patients. J. Gen. Virol. 87: 369-373.

Zou S, Brown EG (1996): Translation of the reovirus M1 gene initiates from the first AUG codon in both infected and transfected cells. Virus Res. 40**:** 75-89.

Appendices

Name	Host	NCBI Accession Number	Genome length	No. of accesssory genes	Name of accessory gene and length of its corresponding protein	
Feline infectious perotinitis virus	Cats	NC_002306	29.4kb	5	ORF3A	71
					ORF3B	71
					ORF3C	345
					ORF7A	101
					ORF7B	206
Porcine respiratory coronavirus	Pigs	DQ811787	27.5 kb	2	ORF3B	205
					ORF7	78
Transmissible gastroenteritis virus	Pigs	DQ811788	28.6 kb	3	ORF3A	71
					ORF3B	244
					ORF7	78
Bat coronavirus CDPHE 15/USA/2006	Bats	NC_022103	28.0 kb	1	ORF3	225
Rousettus bat coronavirus HKU10	Bats	NC_018871	28.5 kb	4	ORF3	218
					ORF7A	81
					ORF7B	153
					ORF7C	76
Miniopterus bat coronavirus1B	Bats	NC_010436	28.5 kb	1	ORF3	219
Miniopterus bat coronavirus1A	Bats	NC_010437	28.3 kb	1	ORF3	219
Miniopterus bat coronavirus HKU8	Bats	NC_010438	28.8 kb	2	ORF3	222
					ORF7	248
Rhinolophus bat coronavirus HKU2	Bats	NC_009988	27.2 kb	2	ORF3	229
					ORF7A	99
Scotophilus bat coronavirus 512	Bats	NC_009657	28.2 kb	1	ORF3	224

Human coronavirus NL63	Humans	NC_005831	27.6 kb	1	ORF3	225
Human coronavirus 229E	Humans	NC_002645	27.3 kb	2	ORF4A	133
					ORF4B	88
Porcine epidemic diarrhea virus	Pigs	NC_003436	28.0 kb	1	ORF3	224

Appendix I, Table 1 enlisting the members of *Alphacoronavirus* with their respective hosts, the length of their genome and number and length of accessory proteins encoded by them.

Name	Host	NCBI Accession Number	Genome length	No. of accessory genes	Name of accessory gene and length of its corresponding protein	
Bovine Coronavirus	Bovine	NC_003045	31.0 kb	6	NS	278
					HE	424
					Vgp05	45
					Vgp06	29
					Vgp07	109
					Vgp10	207
Mouse Hepatitis Virus	Mouse	NC_006852	31.5 kb	5	p30	265
					HE	439
					ORF4	139
					ORF5a	107
					Internal	136
Human Coronavirus HKU1	Humans	NC_006577	29.9 kb	3	HE	386
					ORF4	109
					N2 protein	441
Human Coronavirus OC43	Humans	NC_005147	30.74 kb	5	ORF2a	278
					HE	424
					NS2	109
					NS3	84
					N2	115
SARS Coronavirus	Humans	NC_004718	29.7 kb	8	ORF3a	274
					ORF3b	154
					ORF6	63
					ORF7a	122
					ORF7b	44
					ORF8a	39
					ORF8b	84
					ORF9b	98
Tylonycteris Bat Coronavirus HKU4	Bats	NC_009019	30.3 kb	4	ORF3a	91
					ORF3b	119
					ORF3c	285
					ORF3d	227
Pipistrellus Bat Coronavirus HKU5	Bats	NC_009020	30.5 kb	4	ORF3a	121
					ORF3b	119
					ORF3c	256
					ORF3d	223

MERS Coronavirus	Humans	NC_019843	30.1 kb	5	ORF3a	103
					ORF3b	109
					ORF3c	246
					ORF3d	224
					I Protein	112
Rousettus Bat Coronavirus HKU9	Bats	NC_009021	29.1 kb	3	NS3	220
					ORF7a	185
					ORF7b	149

Appendix I, Table 2 enlisting the members of *Betacoronavirus* with their respective hosts, the length of their genome and number and length of accessory proteins encoded by them.

Name	Host	NCBI Accession Number	Genome length	No. of accessory genes	Name of accessory gene and length of its corresponding protein	
Infectious Bronchitis Virus	Chickens	NC_001451	27.6kb	4	3A	57
					3B	64
					5A	65
					5B	82
Turkey coronavirus	Turkeys	NC_010800	27.6 kb	5	3A	57
					3B	64
					X/4B	94
					5A	65
					5B	82
Beluga whale coronavirus SW1	Beluga whales	NC_010646	31.7 kb	8	5A	138
					5B	172
					5C	175
					6	228
					7	161
					8	59
					9	152
					10	210

Appendix I, Table 3 enlisting the members of *Gammacoronavirus* with their respective hosts, the length of their genome and number and length of accessory proteins encoded by them.

Name	Host	NCBI Accession Number	Genome length	No. of accessory genes	Name of accessory gene and length of its corresponding protein	
Bulbul coronavirus HKU11	Chinese Bulbul	FJ376620	26.5kb	4	NS6	95
					NS7A	123
					NS7B	84
					NS7C	94
Thrush coronavirus HKU12	Grey-backed thrush	NC_011549	26.4 kb	4	NS6	91
					NS7A	123
					NS7B	83
					NS7C	77
Munia coronavirus HKU13	White-rumped munia	NC_011550	26.5kb	4	NS6	108
					NS7A	123
					NS7B	85
					NS7C	93
Porcine coronavirus HKU15	Pig	JQ065042	25.4 kb	2	NS6	94
					NS7	200
White eye coronavirus HKU 16	White eye	JQ065044	26.0 kb	3	NS6	93
					NS7A	222
					NS7B	43
Sparrow coronavirus HKU17	Sparrow	JQ065045	26.1kb	3	NS6	95
					NS7A	144
					NS7B	70
Magpie robin coronavirus HKU18	Magpie robin	JQ065046	26.7kb	4	NS6	96
					NS7A	57
					NS7B	123
					NS7C	84
Night heron coronavirus HKU19	Night heron	JQ065047	26.1kb	3	NS6	92
					NS7A	98
					NS7B	97
Wigeon coronavirus HKU20	Wigeon	JQ065048	26.2kb	5	NS6	90
					NS7A	77
					NS7B	82
					NS7C	88
					NS7D	66
Common moorhen coronavirus HKU21	Common moorhen	JQ065049	26.2kb	4	NS6	81
					NS7A	90
					NS7B	61
					NS7C	138

Appendix I, Table 4 enlisting the members of *Deltacoronavirus* with their respective hosts, the length of their genome and number and length of accessory proteins.

NC_004718	AY282752	DQ412043	DQ412042	AY595412	AY463060
AY485278	AY463059	AY485277	GQ153547	GQ153545	GQ153543
GQ153541	GQ153539	GQ153548	GQ153546	GQ153544	GQ153542
GQ153540	AY310120	AY348314	AY291315	FJ588686	AY313906
AY283797	AY283795	AY283798	AY283796	AY283794	GU553364
GU553365	GU553363	FJ959407	AY279354	AY278487	AY686864
AY278490	AY278489	AY278488	DQ497008	DQ898174	DQ071615
AY304486	AY304488	DQ648856	DQ084200	DQ648857	DQ022305
DQ084199	AY864805	AY686863	AY714217	AY572038	AY572034
AY654624	AY572035	AY390556	AY508724	AY502931	AY502929
AY502927	AY502925	AY502923	AY357075	AY427439	AY502932
AY502930	AY502928	AY502926	AY502924	AY286320	AY357076
AY350750	AY345988	AY345986	AY338174	AY345987	AY338175
AY274119	AY278741	AY278554	DQ182595	AY323977	AY291451
FJ882963	AY772062	EU371563	EU371561	EU371559	EU371564
EU371562	EU371560	DQ640652	AP006560	AP006558	AP006561
AP006559	AP006557	AY545918	AY545916	AY545914	AY545919
AY545917	AY545915	AY613950	AY613948	AY613949	AY613947
AY568539	AY515512	AY559097	AY559095	AY559093	AY559091
AY559089	AY559087	AY559085	AY559083	AY559081	AY559096
AY559094	AY559092	AY559090	AY559088	AY559086	AY559084
AY559082	AY395002	AY395000	AY394998	AY394996	AY394994
AY394992	AY394990	AY394986	AY394978	AY395003	AY395001
AY394999	AY394997	AY394995	AY394993	AY394991	AY394989
AY394987	AY394985	AY394983	AY394979	AY461660	AY304495
AY278491	AY362698	AY362699	AY351680	AY321118	AY297028

Appendix II, Table 1: GenBank accession number of the 156 complete SARS-CoV genomic sequences used in this study.

DQ223042	AB435514	EU004682	DQ911368	EU004663	EU004671
EU854589	EU004660	EU004674	EU004683	EU004664	EU004676
EU004665	EU004672	EF531291	EU004670	DQ223041	EU004677
FJ446720	EU004679	EU004673	EU004668	DQ223043	EU004678
FJ446719	EU004681	EF531290	EU004680		

Appendix II, Table 2: GenBank accession numbers of the 28 complete MNV genomic sequences used in this study.

NC_004718	DQ412043	DQ412042	AY595412	AY463060	AY463059
GQ153547	GQ153545	GQ153543	GQ153541	GQ153539	GQ153548
GQ153546	GQ153544	GQ153542	GQ153540	AY310120	AY291315
FJ588686	GU553364	GU553365	GU553363	AY686864	AY278489
AY278488	DQ497008	DQ898174	DQ071615	DQ084200	DQ022305
DQ084199	AY686863	AY572038	AY572034	AY654624	AY572035
AY390556	AY508724	AY502931	AY502929	AY502927	AY502925
AY502923	AY427439	AY502932	AY502930	AY502928	AY502926
AY502924	AY345988	AY345986	AY345987	AY274119	AY278554
DQ182595	AY323977	AY291451	FJ882963	EU371563	EU371561
EU371559	EU371564	EU371562	EU371560	AP006560	AP006558
AP006561	AP006559	AP006557	KC881007*		

Appendix II, Table 3: GenBank accession numbers of the 70 full-length SARS-CoV genomic sequences used in the mutational model study for the evolution of SARS-CoV overlapping accessory protein 9b study (Chapter 6). * indicates that this sequence was directly obtained from the authors of Ge et al. (2013). Among these, 37 isolates were from human SARS-CoV, 15 from civet SARS-CoV, and 18 (including the newly discovered SL-CoV-WIV1 (Ge et al., 2013)) from bat betacoronaviruses of clade b.

Amino-acid	Codon	Number	Fraction
Ala	GCG	1	0.13
Ala	GCA	4	0.50
Ala	GCT	1	0.13
Ala	GCC	2	0.25
Cys	TGT	0	0.00
Cys	TGC	0	0.00
Asp	GAT	1	0.17
Asp	GAC	5	0.83
Glu	GAG	4	0.80
Glu	GAA	1	0.20
Phe	TTT	0	0.00
Phe	TTC	2	1
Gly	GGG	1	0.33
Gly	GGA	0	0.00
Gly	GGT	1	0.33
Gly	GGC	1	0.33
His	CAT	1	1
His	CAC	0	0.00
Ile	ATA	4	0.80
Ile	ATT	1	0.20
Ile	ATC	0	0.00
Lys	AAG	1	0.33
Lys	AAA	2	0.67
Leu	TTG	2	0.18
Leu	TTA	2	0.18
Leu	CTG	2	0.18

Leu	CTA	1	0.09
Leu	CTT	1	0.09
Leu	CTC	3	0.27
Met	ATG	5	1.00
Asn	AAT	1	0.25
Asn	AAC	3	0.75
Pro	CCG	0	0.00
Pro	CCA	3	0.38
Pro	CCT	0	0.00
Pro	CCC	5	0.63
Gln	CAG	3	0.38
Gln	CAA	5	0.63
Arg	AGG	3	0.60
Arg	AGA	1	0.20
Arg	CGG	0	0.00
Arg	CGA	0	0.00
Arg	CGT	1	0.20
Arg	CGC	0	0.00
Ser	AGT	0	0.00
Ser	AGC	2	0.33
Ser	TCG	0	0.00
Ser	TCA	3	0.50
Ser	TCT	0	0.00
Ser	TCC	1	0.17
Thr	ACG	1	0.13
Thr	ACA	2	0.25
Thr	ACT	1	0.13

Thr	ACC	4	0.50
Val	GTG	6	0.67
Val	GTA	1	0.11
Val	GTT	1	0.11
Val	GTC	1	0.11
Trp	TGG	0	0.00
Tyr	TAT	0	0.00
Tyr	TAC	1	1.00

Appendix III, Table 1: Codon usage values of the 61 amino-acid coding codons in overprinting orf9b gene of SARS-CoV.

Amino-acid	Codon	Number	Fraction
Ala	GCG	25	0.04
Ala	GCA	148	0.29
Ala	GCT	267	0.53
Ala	GCC	71	0.14
Cys	TGT	153	0.65
Cys	TGC	80	0.35
Asp	GAT	255	0.62
Asp	GAC	141	0.38
Glu	GAG	161	0.48
Glu	GAA	186	0.52
Phe	TTT	204	0.60
Phe	TTC	128	0.40
Gly	GGG	14	0.04
Gly	GGA	89	0.18
Gly	GGT	221	0.57
Gly	GGC	95	0.20
His	CAT	106	0.64
His	CAC	54	0.36
Ile	ATA	74	0.17
Ile	ATT	186	0.58
Ile	ATC	83	0.25
Lys	AAG	202	0.50
Lys	AAA	214	0.50
Leu	TTG	124	0.19
Leu	TTA	124	0.16

Leu	CTG	68	0.11
Leu	CTA	67	0.09
Leu	CTT	200	0.30
Leu	CTC	90	0.14
Met	ATG	177	1.00
Asn	AAT	231	0.61
Asn	AAC	134	0.39
Pro	CCG	8	0.03
Pro	CCA	123	0.42
Pro	CCT	118	0.46
Pro	CCC	25	0.09
Gln	CAG	105	0.47
Gln	CAA	129	0.53
Arg	AGG	36	0.13
Arg	AGA	96	0.34
Arg	CGG	3	0.01
Arg	CGA	11	0.05
Arg	CGT	80	0.36
Arg	CGC	33	0.10
Ser	AGT	98	0.21
Ser	AGC	36	0.08
Ser	TCG	13	0.03
Ser	TCA	134	0.26
Ser	TCT	147	0.34
Ser	TCC	30	0.08
Thr	ACG	20	0.03
Thr	ACA	202	0.40

Thr	ACT	207	0.40
Thr	ACC	66	0.17
Val	GTG	113	0.19
Val	GTA	131	0.20
Val	GTT	238	0.46
Val	GTC	98	0.15
Trp	TGG	77	1.00
Tyr	TAT	184	0.56
Tyr	TAC	140	0.44

Appendix III, Table 2 Codon usage values of the 61 amino-acid coding codons in non-overlapping genes of SARS-CoV.

Curriculum Vitae

Personal Details

Name	Aditi Shukla
Date of Birth	19.03.1987
Place of Birth	Ranchi, India

Education

2015 — **PhD**
Graduate School for Computing in Medicine & Life Sciences,
Institute of Biochemistry
University of Lübeck, Germany
Dissertation on "Analysis of overlapping reading frames in viral genomes"

2009 — **Master of Science**
Birla Institute of Technology, Mesra, India
Bioinformatics,
Masters thesis titled "Computer-aided drug discovery for H3N2 influenza virus"

2007 — **Bachelor of Science**
Ranchi University, India
Biotechnology Honors

Scholarship

Ph.D. Scholarship, German Research Foundation (DFG), 2009-2013