Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Virology

Computational methods for inferring location and genealogy of overlapping genes in virus genomes: approaches and applications

Angelo Pavesi

Viruses may evolve to increase the amount of encoded genetic information by means of overlapping genes, which utilize several reading frames. Such overlapping genes may be especially impactful for genomes of small size, often serving a source of novel accessory proteins, some of which play a crucial role in viral pathogenicity or in promoting the systemic spread of virus. Diverse genome-based metrics were proposed to facilitate recognition of overlapping genes that otherwise may be overlooked during genome annotation. They can detect the atypical codon bias associated with the overlap (e.g. a statistically significant reduction in variability at synonymous sites) or other sequence-composition features peculiar to overlapping genes. In this review, I compare nine computational methods, discuss their strengths and limitations, and survey how they were applied to detect candidate overlapping genes in the genome of SARS-CoV-2, the etiological agent of COVID-19 pandemic.

Address
Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, Parco Area delle Scienze 23/A I-43124, Parma, Italy

Corresponding author: Pavesi, Angelo (angelo.pavesi@unipr.it)

Current Opinion in Virology 2022, **52**:1–8

This review comes from a themed issue on **Virus bioinformatics**

Edited by **Alexander Gorbalenya** and **Maria Anisimova**

For complete overview about the section, refer Virus bioinformatics (2021)

Available online 16th November 2021

https://doi.org/10.1016/j.coviro.2021.10.009

Introduction

Overlapping genes, also called dual-coding genes, are genome regions that are translated into two (or more) different reading frames to yield unrelated proteins. They originate by a mechanism of overprinting, in which nucleotide substitutions in a pre-existing non-overlapping open reading frame (ORF) allow the expression of a novel protein from an alternative and originally untranslated ORF, leading to an overlapping-gene arrangement. It is thought that most overlapping genes arose by this mechanism, and that consequently each overlap contains one ancestral frame and one that originated more recently *de novo* [1]. In viruses, in which overlapping genes are abundant [2–4], the mechanism of overprinting is a valuable source of novel proteins [5••], some of which play a crucial role in viral pathogenicity [6–8]. Because freedom to change of each protein in the dual-coding genes is constrained by its counterpart, these genes represent a clear example of adaptive trade-off of the virus evolution under positive selection that promoted the coding expansion of the respective genome regions [9]. The imposed mutual constraints can give rise to proteins with unusual sequence properties [5••], which could be disordered [10,11] or possessing previously unknown 3D structural folds [12•,13•] and mechanisms of action [14].

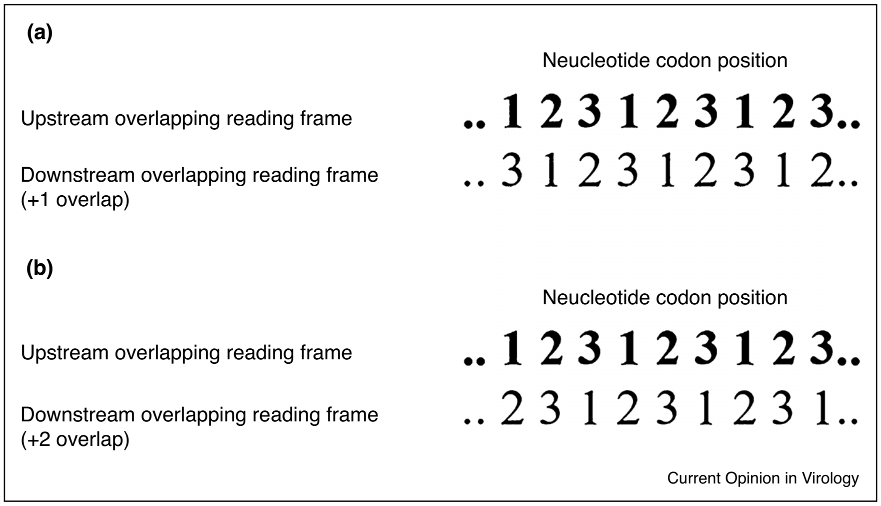
Overlapping genes can be classified broadly into two types: (1) same-strand overlaps, which are transcribed from the same strand of DNA or RNA; (2) different-strand overlaps, which are transcribed from two opposite strands of DNA or RNA. As the great majority of known ORF overlaps are of same-strand type (Figure 1), they were the prime focus of this review. Over the last two decades, the increasing amount of experimentally proven overlapping genes has stimulated several studies (reviewed in Ref. [15]). They also concerned computational methods to detect overlapping genes, some of which were confirmed to encode two unrelated functional proteins during infection [16,17].

In this review, I will first present methods to determine the evolutionary relationship (genealogy) of ORFs in overlapping genes, then methods to detect overlapping genes (Tables 1 and 2), and then survey how they were applied to predict candidate overlapping ORFs in the genome of SARS-CoV-2, the etiological agent of COVID-19 pandemic [18,19]. The term 'ORF' indicates a contiguous stretch of codons, beginning with the most upstream AUG codon, ending with the nearest downstream stop codon, and not interrupted by in-frame stop codons. Candidate ORF means an ORF with evidence for selection pressure indicating that the ORF may thus be beneficial to the virus, while experimental evidence is needed to determine if it is indeed translated into a functional protein.

The genealogy of overlapping genes can be inferred by phylogenetic trees and codon usage

Determining the genealogy of overlapping ORFs means identifying which ORF is ancestral and which one

Figure 1


(a)
Neucleotide codon position
Upstream overlapping reading frame
.. 1 2 3 1 2 3 1 2 3..
Downstream overlapping reading frame (+1 overlap)
.. 3 1 2 3 1 2 3 1 2..
(b)
Neucleotide codon position
Upstream overlapping reading frame
.. 1 2 3 1 2 3 1 2 3..
Downstream overlapping reading frame (+2 overlap)
.. 2 3 1 2 3 1 2 3 1..
Current Opinion in Virology

Orientation of same-strand overlapping genes.
(a) Overlapping gene with the downstream ORF shifted one nucleotide 3′ with respect to the upstream ORF (+1 overlap, known also as −2). It contains three types of codon position (cp): *i)* cp13 in which the first codon position of upstream ORF overlaps the third codon position of downstream ORF; *ii)* cp21 in which the second codon position of upstream ORF overlaps the first codon position of downstream ORF; *iii)* cp32 in which the third codon position of upstream ORF overlaps the second codon position of downstream ORF. **(b)** Overlapping gene with the downstream ORF shifted two nucleotides 3′ with respect to the upstream ORF (+2 overlap, known also as −1). It contains three types of codon position (cp): *i)* cp12 in which the first codon position of upstream ORF overlaps the second codon position of downstream ORF; *ii)* cp23 in which the second codon position of upstream ORF overlaps the third codon position of downstream ORF; *iii)* cp31 in which the third codon position of upstream ORF overlaps the first codon position of downstream ORF. According to the genetic code and on average, a substitution at first codon position causes amino acid change in 95% of cases, at second position in 100% of cases, and at third position in 28% of cases.

originated *de novo*. This can be done by examining their phylogenetic distribution, under the assumption that the protein with the most restricted phylogenetic distribution is encoded by the *de novo* ORF, while that with the widest distribution is encoded by the ancestral ORF [1,5••]. This approach may be inconclusive, if the overlapping ORFs have an identical phylogenetic distribution; in this case, the genealogy can be inferred by applying the codon usage approach [20•]. The later assumes that the ancestral ORF, which has co-evolved over a long period with the other viral genes, must have a codon usage similar to that of other ORFs of the genome. In contrast, the *de novo* ORF has, at birth, a codon usage significantly different from that of the genome, and the constraints imposed by the ancestral ORF might prevent the *de novo* ORF from adopting, later, the typical genome ORF codon usage.

Using basic statistics, such as the Pearson's correlation test, a codon usage analysis of the overlapping genes in bacteriophages ΦX174, α3 and G4 (family *Microviridae*) hypothesized a common ancestor genome having only single-coding genes, whose coding capacity increased over time due to the *de novo* appearance of dual-coding regions in descendants [21]. Using the phylogenetic and codon usage methods, I could predict the ORF genealogy of 46 dual-coding genes from eukaryotic viruses [22•]. By extending the inferred genealogy to the respective orthologs in other viruses, I assembled a dataset of 194 overlapping genes with a known ancestral and *de novo* ORFs [22•]. I will now see methods to detect overlapping genes in viruses. Their features are summarized in Tables 1 and 2.

Methods to detect overlapping genes based on sequence-composition features and codon usage

Development of the sequence-composition methods was favored by the assembly of a first dataset of 80 experimentally proven overlapping genes from 61 distinct virus species [23•]. It was a valuable start point to assemble a much larger dataset [22•], which included the orthologs of each overlapping ORF in other viruses and gathered from the NCBI Viral Genome Database [24]. The size of the sample increased to 319 overlaps, coming from 244 virus species (some viruses contain more than one overlap). Principal component analysis showed that overlapping genes, despite their heterogeneity in length and function, share a common pattern of nucleotide and amino acid composition, which is significantly different from that of the corresponding entire genome complement of non-overlapping genes [22•,23•].

Table 1

Computational methods to detect overlapping genes in viruses

Name of the method	References	Description	Features
SeqComp	[22•]	Detects overlapping genes based on their peculiar nucleotide and amino acid composition.	High sensitivity and low specificity.
CodScr + SeqComp	[25]	Detects overlapping genes based on their peculiar nucleotide and amino acid composition and a statistically significant bias in codon usage.	Good sensitivity and specificity.
Codon test	[26]	Detects overlapping genes on the basis of a length significantly longer than expected by chance; includes a codon-permutation test and a synonymous-mutation test.	High sensitivity for long overlapping genes but intermediate for short overlapping genes. Low specificity.
GOPHIX	[27]	Detects overlapping ORFs on the basis of a significant enrichment in a set of 20 codons that are overrepresented in the protein-coding genes.	Sensitivity and specificity not reported.
Synplot2	[28,37]	Detects overlapping genes by selecting regions with a significantly enhanced conservation at synonymous sites, compared to a null model of neutral evolution.	High sensitivity. Poorly effective for too divergent, or too similar, sequences.
FRESCo	[29]	Detects overlapping genes by selecting regions with an excess of synonymous constraints, under models of neutral and non-neutral evolution.	Good sensitivity and high specificity.
PhyloCSF	[30,31]	Detects overlapping genes by selecting regions evolving under strong protein-coding constraint.	Sensitivity and specificity not reported.
cRegions	[32,33]	Detects overlapping functional elements by identifying regions where the nucleotide sequence is significantly more conserved than expected.	Sensitivity and specificity not reported.
OLGenie	[34,39]	Detects overlapping genes by estimating signs of strong purifying selection.	Intermediate sensitivity and specificity.

Table 2

Further features of the computational methods to detect overlapping genes in viruses

Name of the method and references	Does it require as input single or multiple sequences?	Type of overlap detected	Does it provide a P-value in the results?	Availability
SeqComp [22•]	Single nucleotide sequences	Protein-protein coding	No	Not implemented
CodScr + SeqComp [25]	Single nucleotide sequences	Protein-protein coding	No	Not implemented
Codon test [26]	Single nucleotide sequences	Protein-protein coding	Yes	Script at http://github.com/TimSchlub/Frameshift
GOPHIX [27]	Single nucleotide sequences	Protein-protein coding	No	Not implemented
Synplot2 [28,37]	Multiple homologous sequences	Protein-protein coding or functional RNA element	Yes	Web site at http://www.firthlab.path.cam.ac.uk/virad.html
FRESCo [29]	Multiple homologous sequences	Protein-protein coding or functional RNA element	No	Script at https://www.broadinstitute.org/fresco/fresco
PhyloCSF [30,31]	Multiple homologous sequences	Protein-protein coding	Yes	Script at http://compbio.mit.edu/PhyloCSF
cRegions [32,33]	Multiple homologous sequences	Protein-protein coding or functional RNA element	Yes	Web site at http://bioinfo.ut.ee/cRegions/
OLGenie [34,39]	Multiple homologous sequences	Protein-protein coding	Yes	Script at https://github.com/chasewnelson/OLGenie.

SeqComp

Based on the findings reported in Refs. [22•,23•], I used two statistical methods — the Fisher's linear discriminant analysis (LDA) and the partial least squares-discriminant analysis (PLS-DA) — for distinguishing overlapping genes from non-overlapping genes. Combination of LDA and PLS-DA yielded a sequence-composition method (SeqComp) consisting of two prediction criteria: the LDA score and the PLS-DA score. SeqComp showed a high accuracy, because it correctly classified as overlap 94.2% of 319 overlapping genes and as non-overlap 97.1% of 244 non-overlapping genes of the large dataset assembled in Ref. [22•]. In a subsequent study [25], SeqComp was applied to a dataset of about 4000 spurious overlapping genes, with the aim to calculate its specificity. A spurious overlapping gene consists of a protein-coding

ORF which overlaps with another ORF purely by chance. In analysis of this dataset, SeqComp showed a low specificity, as it classified as true negative only 43.1% of the spurious overlaps [25].

CodScr + SeqComp

To overcome the low specificity of SeqComp, I extended prediction criteria of SeqComp to four and combined this method with a codon scrambling (CodScr) test to produce CodScr + SeqComp method [25]. In addition to the two criteria used by SeqComp [22•], this method included two other criteria. They were obtained from linear discriminant analysis of overlapping genes with known genealogy, which separated ancestral ORF from *de novo* ORF with an accuracy close to 100% [22•]. The CodScr test is based on the assumption that in overlapping genes the use of synonymous codons in the ancestral ORF is significantly biased, to avoid incorporation of premature stop codons in the *de novo* ORF. When applied to a dataset of overlapping genes with known genealogy, CodScr recognized as true positives 166 out of 194 overlaps (sensitivity of 86%). I used this feature as the first prediction criterion in the new method. Under the stringent rule that each overlap must meet all five prediction criteria, CodScr + SeqComp recognized as true positives 157 out of 194 overlaps (sensitivity of 81%; data not shown). When applied to the dataset of about 4000 spurious overlapping genes, CodScr + SeqComp outperformed SeqComp by 42%, as it classified as true negative 85% of the spurious overlaps [25]. This method is yet to be implemented in software.

Codon test

The method developed by Schlub *et al.* [26], here called Codon test, detects candidate overlapping genes in viruses by selecting overlapping ORFs that are significantly longer than expected by chance. The method consists of the codon-permutation and synonymous-mutation tests. In the first test, the expected length of overlapping ORFs is estimated by randomly permuting codon positions in the reference ancestral reading frame. In the other, instead of permuting codon position, the codon order is unchanged and random synonymous mutations are introduced in the reference ancestral reading frame, before measuring ORF lengths in the alternative reading frames. The codon-permutation and synonymous-mutation tests of this method have a rather low specificity (60 and 59%). Another limitation is that the sensitivity of the two tests for ORFs in the range between 100 and 300 nt (65 and 71%) is considerably smaller than for ORFs larger than 300 nt (90 and 95%). The method is available as R script at http://github.com/TimSchlub/Frameshift.

GOFIX

GOFIX (Gene prediction by Open reading Frame Identification using X motifs) is a method that detects ORFs on the basis of a statistically significant enrichment in a set of 20 codons (the motifs of the X circular code) that are overrepresented in the protein-coding genes of a wide range of organisms: bacteria, archaea, eukaryotes, and viruses. By applying GOFIX to coronavirus genomes, Michel *et al.* [27] first found that their structural and accessory genes are significantly enriched in X motifs. In addition, GOFIX predicted two overlapping ORFs conserved in all coronaviruses (ORF9b and ORF9c nested within the nucleocapsid gene) and one overlapping ORF with a restricted phyletic range (ORF3d nested within ORF3a). GOFIX lacks an accessible implementation and its sensitivity and specificity were not reported. Despite these limitations, it is a promising method to characterize potential ORFs, including the overlapping ones, in virus genomes.

Methods to detect overlapping genes based on significant evolutionary constraints at specific sites

The four methods described so far have the advantage that they can be applied to a single viral genome sequence or sequences with few sites of variation that make them especially suitable for detection of newborn or recently born overlapping ORFs. However, when homologous sequences of considerable divergence are available, dual-coding genes can be most readily identified by detecting the atypical pattern of codon bias in two overlapping ORFs induced by the overlap over the course of evolution of the analyzed sequences. This can be done, for example, by identifying genome regions where there is a statistically significant reduction in the degree of variability at synonymous sites. Indeed, this feature should be common to most overlapping genes, because a substitution that is synonymous in one frame is highly likely to be non-synonymous, and not advantageous, in the overlapping ORF (Figure 1). To be effective, application of this approach requires multiple sequences with a substantial range of nucleotide diversity, which are aligned to reveal variation at particular sites.

Synplot2

Synplot2 is a computational tool that analyzes alignments of protein-coding sequences with the aim to identify regions where there is a statistically significant reduction in variability at synonymous sites, which is indicative of an overlapping functional element such as an overlapping gene or a conserved RNA structure [28]. It was one of the first developed and extensively used in the field. When tested on a sample set of 21 representative overlapping genes from 18 virus species, Synplot2 detected 20 overlaps (sensitivity of 95%), all subjected to strong purifying selection. The method may underperform in the following cases when evolutionary signal is weak: *i)* too divergent sequences in which there are too few synonymous positions to assess; *ii)* too similar sequences in which there are too few variations to detect purifying selection, though this drawback can be overcome if sufficiently many low-divergence but non-identical sequences are available. Synplot2 is a web-based, user-friendly method available at http://www.firthlab.path.cam.ac.uk/virad.html.

FRESCo

A computational tool, conceptually similar to Synplot2, was developed by Sealfon *et al.* [29] and called FRESCo (Finding Regions of Excess Synonymous Constraints). FRESCo recovered regions showing a significantly reduced level of synonymous variability in known overlapping genes of well-characterized viral genomes (species *Hepatitis B virus*, *West Nile virus* and *Enterovirus C*). By genome sequence analysis of 30 distinct virus species, FRESCo identified novel regions of excess synonymous constraint in ORF due to an overlapping conserved RNA structure or a putative *de novo* overlapping protein. When applied to dataset of simulated sequences and using a cutoff P-value <0.05, the method did not detect false positives, yielding a specificity of 100%. The site https://www.broadinstitute.org/fresco/fresco provides a script to run FRESCo, sample input files, and instructions on running the method.

PhyloCSF

PhyloCSF (Phylogenetic Codon Substitution Frequencies) is a method that was originally developed to distinguish known protein-coding regions from randomly selected non-coding regions in a multiple whole-genome alignment of 12 *Drosophila* species [30]. PhyloCSF has been widely used for gene annotation and for the discovery of novel protein-coding regions in eukaryotic genomes. For the first time, the method has recently been applied to viral genomes [31], with the aim to detect protein-coding signatures in 44 genome sequences from sarbecoviruses representing intra-species variation (*Sarbecovirus* is a subgenus of *Betacoronavirus* containing only the species *Severe acute respiratory syndrome-related coronavirus*, which includes many viruses of different hosts such as human, pangolin and bat). Interestingly, PhyloCSF detected strong protein-coding signatures for two conserved overlapping ORFs: ORF3c, nested within ORF3a, and ORF9b, nested within the nucleocapsid gene. The performance of the method, in terms of sensitivity or specificity, was not reported. Instructions to download and install PhyloCSF are available at http://compbio.mit.edu/PhyloCSF.

cRegions

This alignment-based method was developed to identify the protein regions in which conservation in the amino acid sequence is caused by an anomalously strong conservation in the nucleotide sequence [32]. The key principle of cRegions is to compare the observed and expected nucleotide frequencies, and calculate three metrics to detect regions where the nucleotide sequence is significantly more conserved than expected. cRegions was able to detect in DNA and RNA viruses functional elements that are under selection, such as splice sites, stem-loops, ribosome frameshifting signals, short overlapping ORFs and other embedded elements with yet unknown function [32]. When applied to human papillomaviruses, cRegions detected two short protein-coding regions (45 and 57 nt) overlapping in the +1 frame the gene encoding the core protein E1 [33]. The cRegions web tool is available at http://bioinfo.ut.ee/cRegions/.

OLGenie

OLGenie, where OLG means OverLapping Gene, is a method that estimates functional constraints on two ORFs in overlapping genes by calculating the ratio of nonsynonymous to synonymous substitutions [34]. It is an extension of an approach designed to evaluate the strength of selection intensities in dual-coding genes, by taking into account the evolutionary constraints acting on interdependent protein-coding sequences encoded in two reading frames [35]. The performance of OLGenie was tested on a sample set of 58 known overlapping genes and a control set of 176 non-overlapping genes, both taken from the database assembled in Ref. [23•]. With a cut-off P value <0.05, OLGenie recognized as true positives 38 overlaps (sensitivity of 66%) and as true negatives 119 non-overlaps (specificity of 68%). Varying P value cut-offs can be used to increase either sensitivity or specificity. OLGenie is available as Perl script at https://github.com/chasewnelson/OLGenie.

Candidate overlapping ORFs detected in SARS-CoV-2

I will now see how these methods were applied to SARS-CoV-2 for detecting candidate overlapping ORFs. Methods for which the results are reported in literature are Codon test, GOFIX, OLGenie, PhyloCSF, Synplot2 and CodScr + SeqComp, with the later application [23•] superseding results obtained with SeqComp [22•]. The aim of this paragraph is assessing to what extent the methods are complementary to each other in a practical setting.

Table 3 shows the six predicted overlapping ORFs, named in accordance to the consensus nomenclature proposed by Jungreis *et al.* [36]. ORF3c and ORF9b, conserved in all sarbecoviruses, were predicted both by methods relying on sequence-composition features (CodScr + SeqComp [25] and GOFIX [27]) and by methods detecting synonymous constraint and protein-coding signature (Synplot2 [37] and PhyloCSF [31]). Ribosome profiling showed that ORF3c and ORF9b are indeed translated during experimental infection of Vero cells [38]. Although conserved in all analyzed sarbecoviruses, ORF9c was predicted only by GOFIX. Antiviral response to SARS-CoV-2 in virus-infected cells was suppressed when the cells were transfected with a ORF9c-encoded plasmid [40], although no evidence

Table 3

List of the candidate overlapping ORFs detected in SARS-CoV-2 using six prediction methods

Candidate overlapping ORF	Length (nt)	Ancestral overlapping gene	Boundaries of the candidate overlapping ORF[a]	Prediction methods
ORF3c	126	ORF3a	25457–25582	CodScr + SeqComp [25], PhyloCSF [31], Synplot2 [37]
ORF3d	174	ORF3a	25524–25697	Codon test [39], CodScr + SeqComp [25], GOFIX [27], OLGenie [39]
ORF9b	294	Nucleocapsid	28284–28577	CodScr + SeqComp [25], GOFIX [27], PhyloCSF [31]
ORF9c	222	Nucleocapsid	28734–28955	GOFIX [27]
ORF-Sh[b]	120	Spike	24051–24170	CodScr + SeqComp [25]
ORF-Mh[b]	180	Membrane	26693–26872	CodScr + SeqComp [25]

[a] Boundaries of the predicted ORFs refer to the reference genome sequence of SARS-CoV-2 (NC_045512.2).
[b] Term ‘h’ stands for hypothetical.

for ORF9c expression during infection was reported. The fourth ORF, ORF3d, showed a restricted phyletic distribution limited to SARS-CoV-2 and Guangxi pangolin-CoVs. It was predicted by four methods: Codon test, CodScr + SeqComp, GOFIX, and OLGenie. Ribosome profiling [38] revealed translation of a shorter isoform of ORF3d, called ORF3d-2 [36], rather its full-length counterpart in infected cells. Like ORF3d, the two last predicted overlapping ORFs, fifth and sixth, showed a narrow phyletic range: ORF-Sh was found in pangolin coronaviruses, Bat-CoV-RaTG13 and SARS-CoV-2, while ORF-Mh was restricted to SARS-CoV-2. Unlike ORF3d, they were predicted by only one method (CodScr + SeqComp). No evidence for translation of these ORFs has (yet) been provided by ribosome profiling.

Finally, SARS-CoV-2 contains also two overlapping ORFs undetected by the above methods (ORF2b and ORF3b, see Figure 1 in Ref. [36]). ORF2b expression was detected by ribosome profiling of infected cells [38]. No such evidence was obtained for ORF3b, although when expressed from a plasmid in Sendai-virus infected cells it was linked to interferon-antagonist function [41]. Since ORF3b is very short (66 nt), it might be recognized using cRegions [32], which seems to be specialized in detecting short overlapping ORFs but yet to be applied to computational analysis of the SARS-CoV-2 genome.

Conclusions and future directions

An increasing number of researchers are becoming aware of the relevance of overlapping genes. Indeed, they also occur in prokaryotes and eukaryotes [42••] and eukaryotic genomes probably contain numerous undetected overlapping genes, as suggested by accumulating experimental evidence [43]. Similarly, virus and bacteriophage genomes contain several candidate overlapping ORFs that would be detectable by computational methods. This review compares nine methods of detecting overlapping genes in viruses, their features and specifics of application. Future directions of research could include a rigorous comparison of sensitivity and specificity and ORF size dependence, by promoting a joint effort of the methods benchmarking using an agreed framework. The dataset of 319 overlaps and that of 194 overlaps with known genealogy assembled in Ref. [22•], and enriched further with newly discovered overlapping genes (e.g. that encoding an essential new protein in astroviruses [44]), could be part of this benchmarking. Computational methods should be web-based and user-friendly, that is the most efficient way to promote their use by the field. This availability, only partially fulfilled at present (Table 2), would facilitate annotation of viral genomes submitted to NCBI database, by including annotations such as ‘candidate overlapping ORF’ or another agreed descriptor. Prediction methods requiring multiple homologous sequences should include plotting start and stop codons in the reading frames alternative to the reference gene, like it was realized in the Synplot2 [28]. This would empower detection of overlapping ORFs that are either fully or partially conserved; in the latter case the possibility of a functional RNA structure could be addressed. A case study of this type could be an overlapping ORF found in GB virus C and nested within the genome region encoding protein NS5A [45]. Finally, the finding that a small set of mammalian overlapping genes follows a sequence-composition bias similar to viral ones [23•] raises the possibility that the presented computational methods may also be applicable to non-viral organisms.

Conflicting and interest statement

Nothing declared.

Acknowledgements

The author expresses deep gratitude to the reviewer David G. Karlin for its valuable feedback. The author is grateful to Chase W. Nelson (Institute for Comparative Genomics, American Museum of Natural History, New York, USA), Rachel Sealfon (Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, USA) and Irwin Jungreis (Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA) for helpful comments and suggestions. This work has benefited from the equipment and framework of the COMP-HUB initiative, funded by the ‘Departments of Excellence’ program of the Italian Ministry for Education, University and Research (MIUR, 2018–2022).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Keese PK, Gibbs A: **Origin of genes: “big bang” or continuous creation?** *Proc Natl Acad Sci U S A* 1992, **89**:9489-9493.

2. Belshaw R, Pybus OG, Rambaut A: **The evolution of genome compression and genomic novelty in RNA viruses**. *Genome Res* 2007, **17**:1496-1504.

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

4. Schlub TE, Holmes EC: **Properties and abundance of overlapping genes in viruses**. *Virus Evol* 2020, **6**:veaa009.

5. Rancurel C, Khosravi M, Dunker KA, Romero PR, Karlin D: •• **Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation**. *J Virol* 2009, **83**:10719-10736

An exhaustive study on the features of the proteins encoded by *de novo* overlapping genes. Almost all are accessory proteins that play a role in viral pathogenicity or spread.

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

7. Vargason JM, Szittya G, Burgyan J, Hall TM: **Size selective recognition of siRNA by an RNA silencing suppressor**. *Cell* 2003, **115**:799-811.

8. Chen W, Calvo PA, Malide D, Gibbs J, Schubert U, Bacik I, Basta S, O’Neill R, Schickli J, Palese P *et al.*: **A novel influenza A virus mitochondrial protein that induces cell death**. *Nat Med* 2001, **7**:1306-1312.

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

10. Lauber C, Kazem S, Kravchenko AA, Feltkamp MC, Gorbalenya AE: **Interspecific adaptation by binary choice at de novo polyomavirus T antigen site through accelerated codon-constrained Val-Ala toggling within an intrinsically disordered region**. *Nucl Acids Res* 2015, **43**:4800-4813.

11. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT *et al.*: **Classification of intrinsically disordered regions and proteins**. *Chem Rev* 2014, **114**:6589-6631.

12. Meier C, Aricescu AR, Assenberg R, Aplin RT, Gilbert RJ, • Grimes JM, Stuart DI: **The crystal structure of ORF-9b, a lipid binding protein from the SARS coronavirus**. *Structure* 2006, **14**:1157-1165

The protein ORF-9b of SARS coronavirus, which shows unusual structural and functional properties, is encoded by an ORF that originated *de novo* within the pre-existing nucleocapsid gene.

13. Baulcombe DC, Molnar A: **Crystal structure of p19 - a universal** • **suppressor of RNA silencing**. *Trends Biochem Sci* 2004, **29**:279-281

The protein p19 of tombusviruses is encoded by an ORF that originated *de novo* within the pre-existing p22 gene. Acting as suppressor of RNA silencing, a mechanism of host defense against infections, p19 is an important pathogenicity factor.

14. Lingel A, Simon B, Izaurralde E, Sattler M: **The structure of the flock house virus B2 protein, a viral suppressor of RNA interference, shows a novel mode of double-stranded RNA recognition**. *EMBO Rep* 2005, **6**:1149-1155.

15. Pavesi A: **Origin, evolution and stability of overlapping genes in viruses: a systematic review**. *Genes* 2021, **12**:809.

16. Jagger BW, Wise HM, Kash JC, Walters KA, Wills NM, Xiao YL, Dunfee RL, Schwartzman LM, Ozinsky A, Bell GL *et al.*: **An overlapping protein-coding region in influenza A virus segment 3 modulates the host response**. *Science* 2012, **337**:199-204.

17. Belhouchet M, Mohd Jaafar F, Firth AE, Grimes JM, Mertens PP, Attoui H: **Detection of a fourth orbivirus non-structural protein**. *PLoS One* 2011, **6**:e25697.

18. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL *et al.*: **A pneumonia outbreak associated with a new coronavirus of probable bat origin**. *Nature* 2020, **579**:265-269.

19. Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, Haagmans BL, Lauber C, Leontovich AM, Neuman BW *et al.*: **The species Severe acute respiratory syndrome-related coronavirus: classifying 2019- nCoV and naming it SARS-CoV-2**. *Nat Microbiol* 2020, **5**:536-544.

20. Pavesi A, Magiorkinis G, Karlin DG: **Viral proteins originated de** • **novo by overprinting can be identified by codon usage: application to the “gene nursery” of Deltaretroviruses**. *PLoS Comput Biol* 2013, **9**:e10031632

This study points out that the pX region of Deltaretroviruses is a hotspot of gene origination. It encodes five genes, at least three of which have originated *de novo* by overprinting.

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

22. Pavesi A: **New insights into the evolutionary features of viral** • **overlapping genes by discriminant analysis**. *Virology* 2020, **546**:51-66

This study shows that overlapping genes, despite their heterogeneity in length and function, share a common pattern of nucleotide and amino acid composition, significantly different from that of non-overlapping genes.

23. Pavesi A, Vianelli A, Chirico N, Bao Y, Blinkova O, Belshaw R, • Firth A, Karlin D: **Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes**. *PLoS One* 2018, **13**:e0202513

This study provides a curated dataset of 80 overlapping genes from 61 virus species. For each overlapping gene, it contains detailed biological information such as type of experimental evidence for expression, mechanism of translation, function of the two encoded proteins, phenotypic effects upon mutation, and bibliography.

24. Brister JR, Ako-Adjei D, Bao Y, Blinkova O: **NCBI viral genomes resource**. *Nucleic Acids Res* 2015, **43**:D571-577.

25. Pavesi A: **Prediction of two novel overlapping ORFs in the genome of SARS-CoV-2**. *Virology* 2021, **562**:149-157.

26. Schlub TE, Buchmann JP, Holmes EC: **A simple method to detect candidate overlapping genes using single genome sequences**. *Mol Biol Evol* 2018, **35**:2572-2581.

27. Michel CJ, Mayer C, Poch O, Thompson JD: **Characterization of accessory genes in coronavirus genomes**. *Virol J* 2020, **17**:131.

28. Firth AE: **Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses**. *Nucleic Acids Res* 2014, **42**:12425-12439.

29. Sealfon RS, Lin MF, Jungreis I, Wolf MY, Kellis M, Sabeti PC: **FRESCo: finding regions of excess synonymous constraint in diverse viruses**. *Genome Biol* 2015, **16**:38.

30. Lin MF, Jungreis I, Kellis M: **PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions**. *Bioinformatics* 2011, **27**:i275-i282.

31. Jungreis I, Sealfon R, Kellis M: **SARS-CoV-2 gene content and COVID-19 mutation by comparing 44 Sarbecovirus genomes**. *Nat Commun* 2021, **12**:2642.

32. Puustusmaa M, Abroi A: **cRegions-a tool for detecting conserved cis-elements in multiple sequence alignment of diverged coding sequences**. *Peer J* 2019, **6**:e6176.

33. Puustusmaa M, Abroi A: **Conservation of the E8 CDS of the E8^E2 protein among mammalian papillomaviruses**. *J Gen Virol* 2016, **97**:2333-2345.

34. Nelson CW, Ardern Z, Wei X: **OLGenie: estimating natural selection to predict functional overlapping genes**. *Mol Biol Evol* 2020, **37**:2440-2449.

35. Wei X, Zhang J: **A simple method for estimating the strength of natural selection on overlapping genes**. *Genome Biol Evol* 2014, **7**:381-390.

36. Jungreis I, Nelson CW, Ardern Z, Finkel Y, Krogan NJ, Sato K, Ziebhur J, Stern-Ginossar N, Pavesi A, Firth AE *et al.*: **Conflicting and ambiguous names of overlapping ORFs in the SARS-CoV-2 genome: a homology-based resolution**. *Virology* 2021, **558**:145-151.

37. Firth AE: **A putative new SARS-CoV protein, 3c, encoded in an ORF overlapping ORF3a**. *J Gen Virol* 2020, **101**:1085-1089.

38. Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Morgenstern D, Yahalom-Ronen Y, Tamir H, Achdout H, Stein D, Israeli O *et al.*: **The coding capacity of SARS-CoV-2**. *Nature* 2021, **589**:125-130.

39. Nelson CW, Ardern Z, Goldberg TL, Meng C, Kuo CH, Ludwig C, Kolokotronis SO, Wei X: **Dynamically evolving novel overlapping gene as a factor in the SARS-CoV-2 pandemic**. *eLife* 2020, **9**: e59633.

40. Dominguez Andres A, Feng Y, Campos AR, Yin J, Yang CC, James B, Murad R, Kim H, Deshpande AJ, Gordon DE *et al.*: **SARS-CoV-2 ORF9c is a membrane-associated protein that suppresses antiviral responses in cells**. *bioRxiv* 2020 http://dx.doi.org/10.1101/2020.08.18.256776. preprint.

41. Konno Y, Kimura I, Uriu K, Fukushi M, Irie T, Kovanagi Y, Sauter D, Gifford RJ, Nakagawa S, Sato K: **SARS-CoV-2 ORF3b is a potent interferon antagonist whose activity is increased by a naturally occurring elongation variant**. *Cell Rep* 2020, **32**:108185.

42. •• Wright BW, Molloy MP, Jaschke PR: **Overlapping genes in natural and engineered genomes**. *Nat Rev Genet* 2021:1-15

An exhaustive review on overlapping genes in prokaryotes, eukaryotes and viruses.

43. Mouilleron H, Delcourt V, Roucou X: **Death of a dogma: eukaryotic mRNAs can code for more than one protein**. *Nucleic Acids Res* 2016, **44**:14-23.

44. Lulla V, Firth AE: **A hidden gene in astroviruses encodes a viroporin**. *Nat Commun* 2020, **11**:4070.

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