Bioinformatics, 35(10), 2019, 1763–1765
doi: 10.1093/bioinformatics/bty851
Advance Access Publication Date: 8 October 2018
Applications Note

OXFORD

Genome analysis

VIRULIGN: fast codon-correct alignment and annotation of viral genomes

Pieter J. K. Libin [1,2,*], Koen Deforche [3], Ana B. Abecasis[4] and Kristof Theys[1]

[1]KU Leuven, Rega Institute for Medical, Laboratorium of Clinical and Evolutionary Virology, 3000 Leuven, Belgium, [2]Artificial Intelligence Lab, Department of Computer Science, Vrije Universiteit Brussel, 1050 Brussels, Belgium, [3]Emweb, 3020 Herent, Belgium and [4]Center for Global Health and Tropical Medicine, Institute for Hygiene and Tropical Medicine, 1349-008 Lisboa, Portugal

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on May 3, 2018; revised on September 24, 2018; editorial decision on September 30, 2018; accepted on October 5, 2018

Abstract

Summary: Virus sequence data are an essential resource for reconstructing spatiotemporal dynamics of viral spread as well as to inform treatment and prevention strategies. However, the potential benefit of these applications critically depends on accurate and correctly annotated alignments of genetically heterogeneous data. VIRULIGN was built for fast codon-correct alignments of large datasets, with standardized and formalized genome annotation and various alignment export formats.

Availability and implementation: VIRULIGN is freely available at https://github.com/rega-cev/virulign as an open source software project.

Contact: pieter.libin@gmail.com

Supplementary information: Supplementary data is available at *Bioinformatics* online.

1 Introduction

Many viral pathogens, in particular RNA viruses, are fast evolving within and between hosts, and markers of adaptation to changing conditions can be detected in their genomes (Lemey *et al.*, 2006). Structural, functional and phenotypic predictions from viral genotypes have fostered advances in drug design, diagnostics and clinical management of viral infections (Houldcroft *et al.*, 2017; Pybus and Rambaut, 2009; Theys *et al.*, 2015). Virus genetic data are also a requisite for inference of evolutionary histories and active epidemiological surveillance (Dellicour *et al.*, 2018; Hadfield *et al.*, 2018; Libin *et al.*, 2017). However, genotype-dependent applications are strongly affected by the quality of underlying sequence alignments.

The process of aligning virus sequences is challenged by their extensive genetic diversity and frequent insertions and deletions, and as a result plethora of alignment software exists with different objectives and applications. Aligners for mapping and assembling sequence reads to study virus populations have significantly advanced in recent years (Posada-Cespedes *et al.*, 2017). Algorithms to align viral consensus or Sanger sequences, resulting in pairwise or multiple alignments, have made less progress over time. Such alignments are however crucial for various aspects of public health and diagnostics.

Multiple sequence alignments (MSAs) of viral genes or genome sequences are often constructed by progressive-iterative approaches such as MAFFT, MUSCLE or Clustal Omega (Edgar, 2004; Katoh and Standley, 2013; Sievers *et al.*, 2011), partly due to their generic applicability and ease of use. These heuristic methods are less capable of mitigating frameshift errors and can be sensitive to noise in sequence data, which is detrimental when protein sequences need to be analyzed in the correct open reading frame (ORF) (e.g. use of codon substitution models in phylogenetics or detection of drug resistance mutations). Alternatively, guidance of the alignment process by a reference sequence can overcome these limitations (Tzou *et al.*, 2017). However, the use of ill-annotated reference sequences hampers the outcome of the alignment. Moreover, inferior sequences in the dataset will have a large impact on the MSA result, and restraining the MSA process by their automatic rejection will further improve the reproducibility and quality of the alignment.

As such, we developed VIRULIGN which is a fast reference-guided and codon-correct alignment and annotation tool for protein coding sequences of closely-related viruses.

2 Related work

In comparison to VIRULIGN, other codon aware alignment softwares are available (e.g. MACSE and TranslatorX) (Abascal *et al.*, 2010; Ranwez *et al.*, 2011), these however do not support to guide the alignment process by reference sequence. HAlign shares VIRULIGN's objective to perform the alignment of large sequence datasets of closely related sequences, but does not focus on codon correct alignments (Zou *et al.*, 2015).

3 Features

VIRULIGN is a cross-platform (GNU/Linux, Unix, MacOS and Windows) and easy-to-use command line application. VIRULIGN can handle large sequence datasets in a computationally efficient manner, as shown experimentally (see Section 5) and through an analysis of the algorithm's computational complexity (see Section 4). Considering a single ORF, VIRULIGN's alignment algorithm is designed for closely related viral genomes with a conserved gene order and corrects the alignment for codon anomalies resulting from single nucleotide alterations. Automated frame shift correction and genome annotation increases the quality of the alignment and reduces the need for manual editing, thereby addressing the need for reproducible research (Peng, 2011).

A codon-correct MSA is essential for evolutionary hypothesis testing and phylogenetic inference using codon substitution models (Shapiro *et al.*, 2005), and for detecting footprints of selective constraints on coding sequence alignments. In addition, the identification of amino acid mutations (including insertions or deletions) associated with drug resistance (HIV-1, Hepatitis C virus, Influenza virus), disease outcome (Hepatitis B virus) or epidemic potential (Ebola virus, Chikungunya virus) are important aspects in the management of infectious diseases.

VIRULIGN enables its users to provide formalized protein annotation of the target CDS, relative to positions within a curated reference genome, through the use of an XML file. This annotation file can be easily defined by the user and VIRULIGN provides predefined annotations for several viral pathogens (see Supplementary Material). The XML file supports the description of a single ORF. In order to handle multi-ORF genomes, multiple annotation files can be specified, to produce distinct alignments for each of the different ORFs, which we demonstrate this in the context of HIV in the Supplementary Material. This feature facilitates genome-wide or protein-specific analyses, and provides virologists with a tool to evaluate and optimize reference sequences in terms of completeness and representativeness (Theys *et al.*, 2017).

VIRULIGN allows to export the computed alignment to various output formats, where different options can be combined to obtain an appropriate alignment representation. Alignments can be exported, either in nucleotide or amino acid alphabet, as FASTA and CSV files, with the latter representing protein positions and mutations as distinct columns.

VIRULIGN is an open-source project (GPLv2 license) written in the C++ programming language. VIRULIGN was previously used in different research areas in infectious diseases (see Supplementary Material for an extensive overview), and can be easily integrated in data management and analysis platforms for viral pathogens.

4 Methods

VIRULIGN attempts to construct an MSA of a set of target sequences $\mathcal{T}$ with respect to the reference sequence r (Figure of the alignment process in Supplementary Material). For each target sequence $t \in \mathcal{T}$, a codon correct pairwise alignment with r is computed. During this procedure, different alignments are generated using the Needleman-Wunsch global alignment algorithm (Needleman and Wunsch, 1970). We will refer to the amino acid representation of the reference sequence r as $AA(r)$.

Firstly, we perform a Needleman-Wunsch nucleotide alignment of r and t, resulting in alignment $\mathcal{A}_{nt}(r,t)$. Secondly, the three ORFs of target sequence t are translated to their respective amino acid representation, and a reference is kept from each of the amino acids to their respective codon. Each of these amino acid sequences is aligned to $AA(r)$ using the Needleman-Wunsch algorithm. From these three alignments, the alignment with the highest alignment score is selected. This amino acid alignment is then converted to a nucleotide alignment $\mathcal{A}_{cc}(r,t)$, by replacing each of the amino acids with their respective nucleotide codon. Thirdly, if $\mathcal{A}_{cc}(r,t)$ and $\mathcal{A}_{nt}(r,t)$ differ, we suspect that a frame-shift has occurred in the target sequence. We then attempt to fix the frame-shift, by detecting the first isolated gap of which the size is not a multiple of three, and replace it by an n nucleotide symbol. Finally, we move again to the second step and the procedure is repeated until no more frame-shifts are detected, or the maximum number of frame-shifts (i.e. a configuration option) has been exceeded. In the latter case, target sequence t is excluded from the MSA, and an error is reported.

This procedure results in a set of codon-correct aligned target sequences, where each of these alignments contains information about possible insertions in the target sequence. This data structure can be exported to a MSA, in a variety of output formats (see Section 3), by iterating over the alignment columns in each of the pairwise alignments.

The way the VIRULIGN algorithm operates, alignment errors will be propagated as frameshift errors. VIRULIGN enables the user to control quality by providing a parameter to bound the number of allowed frameshift corrections.

To derive VIRULIGN's computational complexity, we observe that for each target sequence $t \in \mathcal{T}$, a constant number of Needleman–Wunsch alignments is performed. It is well known that the computational complexity of a Needleman–Wunsch alignment of a sequence tuple (s_1, s_2) is $\mathcal{O}(|s_1| \cdot |s_2|)$ in both space and time (Needleman and Wunsch, 1970). As in VIRULIGN, we consider the reference sequence r and a set of target sequences $\mathcal{T}$, and each target sequence t is aligned to r, the maximal Needleman–Wunsch computational complexity is $\mathcal{O}(|r| \cdot \max(\{|t| t \in \mathcal{T}\}))$. As this applies to all target sequences, VIRULIGN's full computational complexity is $\mathcal{O}(|\mathcal{T}| \cdot |r| \cdot \max(\{|t| t \in \mathcal{T}\}))$.

5 Application and future perspectives

We demonstrate VIRULIGN's abilities by constructing MSAs of real genomic data of Dengue virus (DENV), HIV-1 and Zika virus (ZIKV), which were collected from public databases and also used in studies on viral diversity. Detailed information on the datasets and methods used is available as Supplementary Material.

Firstly, full-length genomes from different genotypes of DENV serotype 1 (DENV-1) ($n = 1433$) were collected from Genbank. This dataset is representative for the DENV-1 worldwide epidemic and was aligned with VIRULIGN, MAFFT, MUSCLE and Clustal Omega. This example shows that, compared to the other tools, VIRULIGN generated an amino acid alignment in the correct ORF

without the need for manual correction, while remaining computationally efficient. Secondly, a selected subset of full-length ZIKV genomes ($n = 19$) was aligned with VIRULIGN using an XML annotation file. The alignment was exported in an amino acid representation to illustrate, in conjunction with other command line utilities, the variability at a glycosylation motif that instigated the effort to correct the ZIKV reference sequence (Theys *et al.*, 2017). Thirdly, we conducted experiments in the context of HIV-1. HIV-1 exhibits three ORFs that together translate the complete set of viral proteins, however, these different reading frames complicate the alignment of the respective CDS. We used a curated set of full-length HIV-1 genomes ($n = 2966$) (Li *et al.*, 2017) that was used to study HIV-1 subtype diversity. This dataset was aligned with VIRULIGN to select the gag poly-protein and identify encoded proteins in an efficient manner. Similar operations can be easily applied to other HIV-1 poly-proteins. As a second example, we used VIRULIGN to align a large HIV-1 dataset ($n = 111\,222$) spanning the reverse transcriptase enzyme, obtained from the curated and public Stanford University HIV Drug Resistance Database (HIVDB). An accurate alignment has significant clinical importance in the context of drug resistance detection. Due to its favorable computational complexity in this context, VIRULIGN performed a better alignment much faster than MAFTT. Through this example, we also demonstrate VIRULIGN's capabilities to exclude erroneous sequences from the alignment (see Supplementary Material). Finally, we demonstrate the strength of VIRULIGN to quickly detect the presence of resistance mutations by reproducing findings from a recent study on HIV drug resistance.

Future developments include a community-driven repository of standardized and curated genome annotations of representative reference sequences and the integration of VIRULIGN in tools for surveillance and genomic epidemiology. Additional areas of interest include the addition of functionalities for multi-ORF alignments, support for non-coding sequences and support for user-defined genetic codes (Taylor *et al.*, 2013). In this work, we focus on virus species with a relatively short genome. Nonetheless, we believe it to be an interesting direction for future work to explore VIRULIGN's potential to align viruses with larger genomes (e.g. orthopoxviruses).

Acknowledgements

Pieter Libin is funded by a doctoral grant of the Research Foundation - Flanders (FWO). This work has been partially funded by the Fundação para a Ciência e Tecnologia (FCT) through funds to GHTM-UID/Multi/04413/2013 and by the MigrantHIV project (PTDC/DTP-EPI/7066/2014). We thank A.-M. Vandamme, S. Imbrechts, L. Cuypers and F. Ferreira for testing the software. We thank the three anonymous reviewers for their comments, as they have helped us to improve the software and manuscript.

Conflict of Interest: none declared.

References

Abascal,F. *et al.* (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.*, **38**(Suppl 2), W7–W13.

Dellicour,S. *et al.* (2018) Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nat. Commun.*, **9**, 2222.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Hadfield,J. *et al.* (2018) Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, **34**, 4121–4123.

Houldcroft,C.J. *et al.* (2017) Clinical and biological insights from viral genome sequencing. *Nat. Rev. Microbiol.*, **15**, 183–192.

Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

Li,G. *et al.* (2015) An integrated map of HIV genome-wide variation from a population perspective. *Retrovirology*, **12**, 18.

Lemey,P. *et al.* (2006) HIV evolutionary dynamics within and among hosts. *AIDS Rev.*, **8**, 125–140.

Libin,P. *et al.* (2017) PhyloGeoTool: interactively exploring large phylogenies in an epidemiological context. *Bioinformatics*, **33**, 3993–3995.

Needleman,S. and Wunsch,C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Peng,R.D. (2011) Reproducible research in computational science. *Science*, **334**, 1226–1227.

Posada-Cespedes,S. *et al.* (2017) Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus Research*, **239**, 17–32.

Pybus,O. and Rambaut,A. (2009) Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.*, **10**, 540 EP.

Ranwez,V. *et al.* (2011) MACSE: multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS One*, **9**, e22594.

Shapiro,B. *et al.* (2005) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol. Biol. Evol.*, **23**, 7–9.

Sievers,F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

Taylor,D.J. *et al.* (2013) Virus-host co-evolution under a modified nuclear genetic code. *PeerJ.*, **1**, e50.

Theys,K. *et al.* (2015) Discordant predictions of residual activity could impact dolutegravir prescription upon raltegravir failure. *J. Clin. Virol.*, **70**, 120–127.

Theys,K. *et al.* (2017) Zika genomics urgently need standardized and curated reference sequences. *PLoS Pathog.*, **13**, e1006528.

Tzou,P.L. *et al.* (2017) NucAmino: a nucleotide to amino acid alignment optimized for virus gene sequences. *BMC Bioinformatics*, **18**, 138.

Zou,Q. *et al.* (2015) HAlign: fast multiple similar DNA/RNA sequence alignment based on the centre star strategy. *Bioinformatics*, **15**, 2475–2481.