BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 21 2011, pages 3074–3075
doi:10.1093/bioinformatics/btr519

Data and text mining
Advanced Access publication September 11, 2011

Metavir: a web server dedicated to virome analysis

Simon Roux[1,2,*], Michaël Faubladier[1,2], Antoine Mahul[3], Nils Paulhe[1,2], Aurélien Bernard[1,2], Didier Debroas[1,2] and François Enault[1,2]

[1]Laboratoire Microorganismes: Génome et Environnement, Clermont Université, Université Blaise Pascal, BP 10448, F-63000 Clermont-Ferrand, [2]CNRS, UMR 6023, LMGE, F-63177 Aubiere and [3]Centre Régional de Ressources Informatiques, Clermont Université, Université Blaise Pascal, Clermont-Ferrand, France

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Metavir is a web server dedicated to the analysis of viral metagenomes (viromes). In addition to classical approaches for analyzing metagenomes (general sequence characteristics, taxonomic composition), new tools developed specifically for viral sequence analysis make it possible to: (i) explore viral diversity through automatically constructed phylogenies for selected marker genes, (ii) estimate gene richness through rarefaction curves and (iii) perform cross-comparison against other viromes using sequence similarities. Metavir is thus unique as a platform that allows a comprehensive virome analysis.

Availability: Metavir is freely available online at: http://metavir-meb.univ-bpclermont.fr

Contact: simon.roux@univ-bpclermont.fr

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on May 6, 2011; revised on July 29, 2011; accepted on August 29, 2011

1 INTRODUCTION

Describing environmental viral communities is a major challenge for environmental microbiology. Viruses are known to intervene in a broad spectrum of processes spanning population regulation, horizontal gene transfer and major biogeochemical cycles (Suttle, 2007). However, the study of environmental viral communities is made difficult by the fact that only a tiny fraction of their hosts has been cultivated. Moreover, as no single gene is common to all viral genomes, environmental viral communities cannot be studied via approaches based on ribosomal RNA sequencing (Edwards and Rohwer, 2005). One way around these limitations is to directly sequence the viral communities. Such metagenomic approaches can provide insights into the viral diversity of environments of interests and are progressively gaining wider uses (Allen and Wilson, 2008). Existing bioinformatics tools implemented on web servers dedicated to metagenome analyses are not specific to particular biological entities (Meyer *et al.*, 2008; Seshadri *et al.,* 2007). Yet, viral metagenomes (viromes) are by nature significantly different from bacterial metagenomes. Indeed, very little is known about viral communities and the strong sequence divergence and broad gene richness found in viromes indicate a tremendous genetic diversity (Kristensen *et al.,* 2009). Yet this diversity remains very poorly characterized as the databases lack annotated viral sequences. Most virome sequences (i.e. reads) therefore differ from previously described sequences (Edwards and Rohwer, 2005). However, direct comparisons of virome sequences from disparate locations often exhibit significant overlap, indicating that, although our knowledge on viral genes remains sparse, the same genes are seen everywhere (Polson *et al.* 2011). The unknown fraction of the reads (between 65% and 95%) usually left aside in classical metagenomic analyses represents extremely valuable data for virome analysis. Furthermore, different marker genes describing the different viral families are needed to further describe viral diversity and none of these markers are available on existing ‘generalist’ web servers.

To our knowledge, PHACCS (Phage Communities from Contig Spectrum; Angly *et al.*, 2005) is the only tool designed for viral community sequencing, but it focuses on assessing the structure of uncultured viral communities in terms of ecological parameters.

Here we present Metavir, an interactive web server that performs a comprehensive analysis of viromes. Classical analyses are available (sequence characteristics, taxonomic composition). Marker genes selected for each major viral families can be used in a custom-designed procedure to perform phylogenetic analysis on virome reads. The automatically generated phylogenetic trees allow biologists to explore the viral diversity in a deep and precise manner. Finally, specific tools have been developed to efficiently deal with the vast unknown fraction. The gene richness of a virome can be assessed and compared to other viromes, and viromes can also be compared in terms of sequence similarity.

2 METHODS

Metavir provides users with a suite of tools inside a private environment for analyzing viromes. After a registration step, the user can submit viromes as fasta files. Metavir separates virome analysis into four major tools, as illustrated in the Supplementary Material using the Sargasso Sea virome (Angly *et al.*, 2009).

2.1 Virome composition

Virome composition is assessed using the GAAS tool (Angly *et al.*, 2009). Virome reads are compared to complete viral genomes from the Refseq database, and taxonomic affiliation results are normalized by genome length in order to estimate the number of viral particles for each viral species in the initial sample (Supplementary Fig. S1).

[*]To whom correspondence should be addressed.

2.2 Automatic phylogenies for marker genes

A procedure has been developed to insert metagenomic reads in phylogenetic trees containing reference sequences for chosen marker genes. This is the first automatic phylogeny generation procedure available for virome sequences. Phylogenies are of great utility to virome research due to the tremendous diversity observed in viral sequences and the lack of representative sequences in the databases. If enough reads are homologous to the marker gene, phylogenies can be generated from 100 bp reads but the procedure provides even better results with 400 bp reads commonly generated today by NGS tools such as 454 TITANIUM.

For each marker available, reference sequences have been retrieved from the PFAM database and aligned using MUSCLE (Edgar, 2004). A BLASTx is computed to detect potential homologous sequences in the virome, and all metagenomic reads having a BLAST hit against one of the reference sequences with an E-value $< 10^{-3}$ are gathered. These sequences are then compared to NR (BLASTx), and excluded from the analysis if their best BLAST hit does not correspond to the studied marker. The remaining reads are assembled using Cap3 (Huang and Madan, 1999) (98% identity on 35 bp) to be able to work with longer sequences. These parameters should only group sequences from the same virotype (Angly *et al*., 2005). These sequences are translated into protein sequences and then aligned against the reference alignment via a HMM profile using HMMER (Eddy, 1998). In order to generate trees containing several metagenomic sequences, alignment bounds for each metagenomic sequence are collected and used to define multiple subalignments. Alignments are cleaned using Gblocks (Talavera and Castresana, 2007) and used to generate phylogenetic trees with 100 bootstraps using PhyML (Guindon *et al*., 2009). Finally, the tree is rooted and monophyletic groups are highlighted via Scriptree (Chevenet *et al*., 2010) (Supplementary Fig. S2). This analysis is already available for the main viral families through different marker genes, such as VP1 for *Microviridae*, or TerL for *Caudovirales.* Users can request specific marker genes using a form on the website.

2.3 Virome comparison

In order to compare viromes in their entirety rather than only their small known fraction, a qualitative comparison of viromes based on sequence similarity (tBLASTx comparison) is computed as described in a previous work on bacterial metagenomes (Martín-Cuadrado *et al*., 2007). Virome samples of 50 000 sequences of 100 bp are used in order to have comparable results. Each sample is compared to every other sample using tBLASTx. A similarity score between virome A and virome B is then computed as the sum of best BLAST hit scores of virome A reads against virome B reads. Finally, the resulting score matrix (i.e. similarity scores for all virome pairs) is used to cluster viromes using R software and the pvclust package, working with default parameters and 100 bootstraps (Suzuki and Shimodaira, 2006). Users can choose to compare any virome subsets (Supplementary Fig. S3).

2.4 Rarefaction curves

Here again utilizing the whole virome rather than only the reads identified by BLAST, the rarefaction curves are computed to assess the gene richness of the viromes (Raes and Bork, 2008). Viromes can then be compared using this view of the genetic diversity within a viral community. Clusters are computed with Uclust (Edgar, 2010) and three thresholds are proposed: 75, 90 and 98%. A Perl script counts the number of different clusters generated for a given number of input sequences, in order to plot rarefaction curves. These rarefaction curves are computed on whole viromes (Supplementary Fig. S4), which makes it possible to determine whether the entire gene pool is sampled in the virome (in which case the rarefaction curve would level off), and on virome samples (50 000 sequences of 100 bp), in order to compare gene richness between different ecosystems (Supplementary Fig. S5). Users can select any virome subset to be plotted into rarefaction curves. Curves are dynamically generated with JS Charts.

3 WEB INTERFACE AND IMPLEMENTATION

A set of previously published viromes has already been included in Metavir as public projects available for any user (registered or not). Private projects are restricted to the user who uploaded the project, until such time as the user decides to make it public. Metavir computations are distributed on a cluster allowing multiple parallel runs (40 CPU). A full analysis of a large-sized virome (600 000 reads of 400 pb) would take a few days in total.

Funding: This work was supported by the French ‘Defence Procurement Agency’ (DGA).

Conflict of Interest: none declared.

REFERENCES

Allen,M.J. and Wilson,W.H. (2008) Aquatic virus diversity accessed through omic techniques: a route map to function. *Curr. Opin. Microbiol.,* **11**, 226–232.

Angly,F. *et al*. (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics*, **6**, 41.

Angly,F. *et al*. (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput. Biol*, **5**, e1000593.

Chevenet,F. *et al*. (2010) ScripTree: scripting phylogenetic graphics. *Bioinformatics*, **26**, 1125–1126.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Edwards,R.A. and Rohwer,F. (2005) Viral metagenomics. *Nat. Rev. Micro.*, **3**, 504–510.

Guindon,S. *et al*. (2009) Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.*, **537**, 113–137.

Huang,X. and Madan,A. (1999) CAP3: a DNA Sequence Assembly Program. *Genome Res.*, **9**, 868–877.

Kristensen,D.M. *et al.* (2010) New dimensions of the virus world discovered through metagenomics. *Trends Microbiol.,* **18**, 11–19.

Martín-Cuadrado,A.-B. *et al*. (2007) Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS ONE*, **2**, e914.

Meyer,F. *et al*. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Polson,S.W. *et al*. (2011) Unraveling the viral tapestry (from inside the capsid out). *ISME J.*, **5**, 165–168.

Raes,J. and Bork,P. (2008) Molecular eco-systems biology: towards an understanding of community function. *Nat. Rev. Microbiol.,* **6**, 693–639.

Suttle,C.A. (2007) Marine viruses—major players in the global ecosystem. *Nat. Rev. Micro.*, **5**, 801–812.

Suzuki,R. and Shimodaira,H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, **22**, 1540–1542.

Seshadri,R. *et al.* (2007) CAMERA: a community resource for metagenomics. *PLoS Biol.,* 2007, **5**, e75.

Talavera,G. and Castresana,J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.*, **56**, 564–577.