Bioinformatics, 32(13), 2016, 2041–2043
doi: 10.1093/bioinformatics/btw110
Advance Access Publication Date: 26 February 2016
Applications Note

Sequence analysis

Computational clustering for viral reference proteomes

Chuming Chen[1,*], Hongzhan Huang[1], Raja Mazumder[2], Darren A. Natale[3], Peter B. McGarvey[3], Jian Zhang[3], Shawn W. Polson[1], Yuqi Wang[1], Cathy H. Wu[1,3] and UniProt Consortium[1,3,4,5]

[1]Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE 19711, USA, [2]Department of Biochemistry and Molecular Medicine, The George Washington University, Washington, DC 20037, USA, [3]Protein Information Resource, Georgetown University Medical Center, Washington, DC 20007, USA, [4]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [5]Swiss Institute of Bioinformatics, Centre Medical Universitaire, Geneva 4 1211, Switzerland

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on December 2, 2015; revised on January 20, 2016; accepted on February 21, 2016

Abstract

Motivation: The enormous number of redundant sequenced genomes has hindered efforts to analyze and functionally annotate proteins. As the taxonomy of viruses is not uniformly defined, viral proteomes pose special challenges in this regard. Grouping viruses based on the similarity of their proteins at proteome scale can normalize against potential taxonomic nomenclature anomalies.

Results: We present Viral Reference Proteomes (Viral RPs), which are computed from complete virus proteomes within UniProtKB. Viral RPs based on 95, 75, 55, 35 and 15% co-membership in proteome similarity based clusters are provided. Comparison of our computational Viral RPs with UniProt's curator-selected Reference Proteomes indicates that the two sets are consistent and complementary. Furthermore, each Viral RP represents a cluster of virus proteomes that was consistent with virus or host taxonomy. We provide BLASTP search and FTP download of Viral RP protein sequences, and a browser to facilitate the visualization of Viral RPs.

Availability and implementation: http://proteininformationresource.org/rps/viruses/

Contact: chenc@udel.edu

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

1 Introduction

There are existing methods for reducing protein sequence space. NCBI-nr (NCBI Resource Coordinators, 2015) clusters identical proteins, while UniRefs (Suzek *et al.*, 2015) provide clustered protein sequences at multiple similarity levels. NCBI-RefSeq (Tatusova *et al.*, 2015) and UniProt's Reference Proteomes (The UniProt Consortium, 2015) provide users with the ability to perform analyses or create protein families using a limited set of complete proteomes selected by curators. However, such approaches are difficult to scale up and do not show the relationships between organisms.

Previously, we created Representative Proteomes (RPs) (Chen *et al.*, 2011) by clustering Archaea, Bacteria and Eukaryota proteomes into related groups and selecting a representative organism from each group. Compared to UniRefs clustering, which is based on the similarity of individual proteins, RPs clustering is based on the similarity of whole proteomes. These computationally determined RPs are now included in the UniProt Reference Proteome set (The UniProt Consortium, 2015). However, viral proteomes posed special challenges, as the taxonomy for viruses is not uniformly defined; potentially there are taxonomic nomenclature variations resulting from naming more distantly related viruses as if they were the same genus or by naming closely related viruses as if they were in different taxonomic clades. To address those challenges we clustered virus proteomes based on the similarity of their overall protein sequences in a proteome scale using algorithms that can reliably and quickly calculate a hierarchical set of Viral Reference Proteomes (Viral RPs) at different co-membership cut-offs. Specifically, the clustering of the virus proteomes starts at very high similarity level in order to identify the proteomes that are very similar to other proteomes or proteomes that are an identical subset of other proteomes (sub-proteomes). In addition, we incorporate the UniProtKB annotation score (http://www.uniprot.org/help/annotation_score) (The UniProt Consortium, 2015) into our Proteome Priority Score (PPS) (Chen *et al.*, 2011).

2 Methods

Viral RPs are updated monthly using UniProtKB complete proteomes and UniRef50. Each proteome used in the calculation is rendered approximately 'one entry per gene'. We take all Swiss-Prot entries for a given gene in each proteome, sort the TrEMBL entries by sequence length and retain only the entry with the longest sequence for any gene not already represented by a Swiss-Prot entry. The procedure to generate Viral RPs includes: (i) Compute pair-wise co-membership for all Virus proteomes. (ii) Cluster Virus proteomes into Representative Proteome Groups (RPGs). (iii) Select Reference proteomes from each RPG.

2.1 Co-membership of two proteomes

The relatedness of two proteomes is measured by their co-membership (X) in UniRef50 clusters and is defined as:

$$X = \frac{2 \times 100 \times N_{ab}}{N_a + N_b}(\text{cut-offs} < 90\%);\ \text{otherwise},$$

$$X = \frac{100 \times N_{ab}}{N_{\mathrm{a}}},\ N_a < N_b$$

where N_a is the number of UniRef50 clusters containing a protein from proteome A. N_b is the number of UniRef50 clusters containing a protein from proteome B. N_{ab} is the number of UniRef50 clusters containing a protein from both proteomes A and B. Given a co-membership cut-off (CMC), if $X \geq \text{CMC}$, two proteomes are grouped together.

2.2 Representative proteome group (RPG)

The co-membership (X) of each pair of proteomes is calculated for all virus complete proteomes. The mean co-membership for a proteome is the average X value between a proteome and all other proteomes. RPGs are constructed as follows: (i) proteomes are ranked according to their average X values. (ii) The top proteome is taken as the seed for a new RPG. For the rest of the proteomes in the ranked list, if $X \geq \text{CMC}$ with the seed, they are added to the newly formed group and then removed from the ranked list. (iii) Re-calculate mean co-membership for the proteomes in the list. (iv) Repeat steps i, ii and iii. (v) Annotate the virus proteomes in each RPG with their taxonomic group and host information derived from UniProtKB and NCBI Genome databases.

RPGs were calculated for five CMC cut-off levels (95, 75, 55, 35 and 15%) hierarchically using a top-down approach to ensure the proteomes grouped together stay together even at a lower CMC.

2.3 Select representative from RPG

Proteomes in each RPG are scored to facilitate representative selection for the group based on: (i) number of unique PubMed references listed in the UniProtKB entries in the proteome, excluding large-scale analysis (PMID). (ii) Mean UniProt Annotation Score of UniProtKB protein entries in the proteome (MeanAS). (iii) Number of UniProtKB protein entries in the proteome (Entry). The UniProt reference proteome (RefP) and the previously released representative proteome (PRP) are given higher weights for stability. The Proteome Priority Score (PPS) for representative selection is calculated as follows: PPS = 10000 × (RefP) + 8000 × (PRP) + 1000 × (PMID) + 100 × (MeanAS) + 1 × (Entry).

3 Results and discussions

One objective in producing a Viral RP set is to reduce the protein sequence space while preserving both sequence representation and annotation. Table 1 shows that low CMCs tend to greatly reduce sequence space at the cost of producing RPGs that may contain multiple genera in one RPG. High CMCs have a low sequence space reduction, and are more likely to split species from a given genus into multiple RPGs. Nonetheless, even at 95% CMC, only ~5% virus proteomes from the same species are split among multiple RPGs. Close examination reveals that a majority of these are bacteriophages, where enormous diversity exists even within those found in the same bacterial host (Grose and Casjens, 2014). As expected, at all cut-offs the taxonomic consistency of RPGs rises with taxonomic rank; that is, successively fewer RPGs contain viruses from multiple taxa as one traverses from species (not shown) to genus to family. This is also observed as one traverses the taxonomic ranks of virus hosts (Supplementary Fig. S1). The algorithmically selected viral RPs are mostly consistent with the ones selected manually by UniProt curators (Supplementary Table S2).

We developed a web site to disseminate Viral RPs and related data. Users can download pre-selected Viral RP sequences or custom build sequence files with respect to taxonomic groups and CMCs. We developed a Viral RPs browser that shows the proteomes at different co-membership cut-offs and provides taxonomic and host

Table 1. Summary statistics of Viral RPGs

CMC (%)	# RPG	# RefP seqs versus # Seqs of proteomes (%)	# Proteome Reduction (%)	Species Split (%)	Multiple Genus RPG (%)	Multiple Family RPG (%)
95	2299	84.75	30.58	4.79	0.24	0.09
75	1683	56.28	49.18	1.98	0.62	0.26
55	1354	43.74	59.13	1.21	0.78	0.49
35	1147	36.45	65.37	0.73	1.23	0.68
15	945	28.33	71.47	0.47	3.89	0.95

Based on UniProtKB release 2015_10; # of Seqs of proteomes: 219525; # of proteomes: 3312; CMC: Co-membership cut-off; RPG: Representative Proteome Group.

information. A BLASTP search is provided against Viral RP sequences.

A standard set of Viral RPs should facilitate functional annotation by decreasing the redundancy of similarity search results, thereby aiding the identification of homologs, protein family classification and comparative genomic and proteomic analyses. RPGs can also be used to identify errors in proteome annotations. For example, identical species are expected to be in the same cluster at 95%. When they are not it is often due to an annotation error.

Taxonomic assignment among viruses is usually derived from phylogenetic analyses. Nonetheless, grouping on the basis of shared proteome features has been suggested as a more accurate view of viral relatedness for phages because of their high recombination rate (Lima-Mendez *et al.*, 2007; Rohwer and Edwards, 2002). By taking advantage of the tunable CMC resolution, viral RPGs have the potential to serve as a tool not just for grouping viruses into broad categories, but also for differentiating closely related strains.

The inability to discern a virus' host range from its genome sequence has also represented a significant problem in interpretation of viral metagenomic data. The host range of viruses observed in RPGs with high CMC tends to be quite narrow, with about 96% of RPG95s sharing a species-level host among all members, while the same viruses will often occur in RPGs with broadening host ranges at successive higher CMC (Supplementary Fig. S1). By exploiting this trend, the known host range within a given RPG may be used to provide clues toward the host range of unknown viruses in the same group.

Funding

The project is supported by the National Institutes of Health (NIH) under the UniProt Consortium grant (Grant No. U41HG007822) and the Delaware INBRE program (Grant No. P20GM103446) and institutional resources of the Center for Bioinformatics and Computational Biology at University of Delaware.

Conflict of Interest: none declared.

References

Chen,C. *et al.* (2011) Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. *PLoS ONE*, **6**, e18910.

Grose,J.H. and Casjens,S.R. (2014) Understanding the enormous diversity of bacteriophages: the tailed phages that infect the bacterial family Enterobacteriaceae. *Virology*, **468–470**, 421–443.

Lima-Mendez,G. *et al.* (2007) Analysis of the phage sequence space: the benefit of structured information. *Virology*, **365**, 241–249.

NCBI Resource Coordinators (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **43**, D6–17.

Rohwer,F. and Edwards,R. (2002) The phage proteomic tree: a genome-based taxonomy for phage. *J. Bacteriol.*, **184**, 4529–4535.

Suzek,B.E. *et al.* (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **6**, 926–932.

Tatusova,T. *et al.* (2015) Update on RefSeq microbial genomes resources. *Nucleic Acids Res.*, **43**, D599–D605.

The UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.