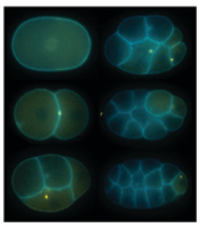
Communicative & Integrative BIOLOGY



Communicative & Integrative Biology

Taylor & Francis
Taylor & Francis Group

ISSN: (Print) 1942-0889 (Online) Journal homepage: https://www.tandfonline.com/loi/kcib20

A network perspective on the virus world

Jaime Iranzo, Mart Krupovic & Eugene V. Koonin

To cite this article: Jaime Iranzo, Mart Krupovic & Eugene V. Koonin (2017) A network perspective on the virus world, Communicative & Integrative Biology, 10:2, e1296614, DOI: 10.1080/19420889.2017.1296614

To link to this article: https://doi.org/10.1080/19420889.2017.1296614

The article not subject to US copyright law

Accepted author version posted online: 23 Feb 2017.
Published online: 23 Feb 2017.

Submit your article to this journal

Article views: 886

View Crossmark data

Citing articles: 10 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=kcib20

COMMUNICATIVE & INTEGRATIVE BIOLOGY
2017, VOL. 10, NO. 2, e1296614 (4 pages)
http://dx.doi.org/10.1080/19420889.2017.1296614

Taylor & Francis
Taylor & Francis Group

SHORT COMMUNICATION

OPEN ACCESS

A network perspective on the virus world

Jaime Iranzo[a], Mart Krupovic[b], and Eugene V. Koonin[a]

[a]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, USA; [b]Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extrêmophiles, Paris, France

ABSTRACT

Viral evolution is characterized by high rates of horizontal gene transfer and fast sequence divergence. Furthermore, there are no universal genes shared by all viruses. As a result, distant relationships among viruses are better represented by a network than by a tree. Here we discuss 3 network representations of the virus world with decreasing levels of complexity, from a multilayer network that integrates sequence conservation and patterns of gene sharing to a classic genome similarity network. As new tools for network analysis are developed, we expect that novel insights into virus evolution will result from the study of more complex representations of the virus world.

ARTICLE HISTORY
Received 18 January 2017
Revised 10 February 2017
Accepted 13 February 2017

KEYWORDS
bipartite network; gene sharing network; multilayer network; phylogenomics; viral evolution; viral taxonomy

Since *The Origin of Species* was published, the idea that all extant and past forms of life can be organized as a Tree of Life (TOL) has become quintessential to evolutionary biology. More than 150 years later, in the wake of the genomic revolution, the TOL remains a valid approximation, as long as horizontal gene transfer (HGT) among prokaryotes and viruses does not completely blur the tree-like pattern that arises from vertical descent with modification.[1] While HGT sets a fundamental limit to the TOL concept, high degrees of sequence divergence impose a technical limitation to the construction of deep phylogenetic trees. Because fast divergence and intense horizontal transfer are 2 main characteristics of viral evolution, the reconstruction of large-scale viral phylogenies poses a major technical and fundamental challenge to the interpretation of the virosphere as a tree. Furthermore, there are no universal genes shared by all or even most groups of viruses, which restricts phylogenetic analyses to discrete assemblages of more closely related viruses, thereby fragmenting and blurring the global understanding of the evolutionary relationships in the viral world. Instead, the evolutionary relationships among viruses can be more precisely represented as a network of gene sharing.[2]

Among many possible network representations,[3] a highly informative description of the virus world is provided by a network with 2 layers (Fig. 1A). The gene layer consists of a sequence similarity network, with nodes representing viral genes and edges connecting pairs of homologous genes with a weight proportional to their sequence similarity. The second layer represents the viral genomes; nodes in the genome layer do not connect with each other, but rather to nodes from the gene layer: each genome node is connected to the genes that belong to that genome. A simpler representation of this 2-layer network can be obtained by aggregating the nodes from the gene layer into groups of homologous genes. Such groups appear in the gene layer as modules, i.e. sets of nodes that are much more densely connected with each other than with the rest of the network. Indeed, some popular methods to identify sets of orthologous or, more generally, homologous genes work by applying a module detection algorithm to a sequence similarity network.[4,5] Once genes are grouped into families of homologs, a bipartite network is obtained by connecting genome nodes and gene family nodes whenever a gene family is present in a genome (Fig. 1B). Compared with the 2-layer network, the bipartite network lacks the former's precise information on sequence similarity, which could be used to reconstruct single-gene phylogenies, but keeps the essential information on which gene families are shared by which genomes. A further simplification results from projecting the bipartite network into a genome similarity network (Fig. 1C). There are

CONTACT Eugene V. Koonin koonin@ncbi.nlm.nih.gov National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

Short Communication to: Iranzo J, Krupovic M, & Koonin EV. The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. MBio 2016; 7(4):e00978-16; http://dx.doi.org/10.1128/mBio.00978-16

The article not subject to US copyright law.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

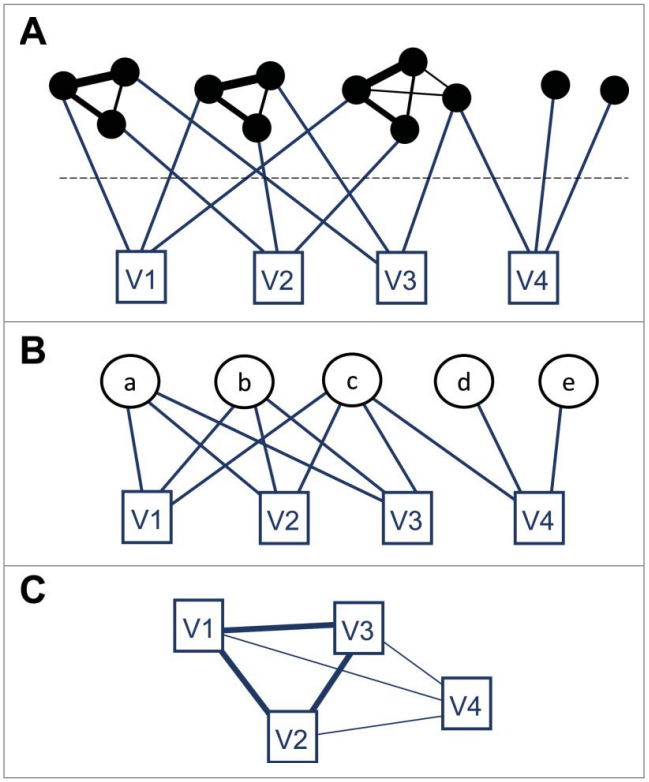


Figure 1. Three network representations of a toy virus world composed of 4 viral genomes (squares) and 12 genes (black circles) that belong to 5 gene families (white circles). (A) Two-layer network, with the gene layer on top and the genome layer at the bottom. Black edges of different thickness indicate the similarity between sequences in the gene layer. (B) Bipartite network, which results from clustering groups of homologous genes in the gene layer into gene family nodes. (C) Genome similarity network; the thickness of the links is proportional to the number of shared gene families.

multiple ways to obtain a genome similarity network from a bipartite gene sharing network. The main step is defining a measure of similarity between genomes, such as the number of shared genes, the fraction of shared genes, or the probability that such number of shared genes occurs by chance. The network is then readily obtained by connecting pairs of genomes with edges, whose weights are proportional to the similarity between the genomes.

Within the network framework, module detection algorithms[6,7] have become a useful tool to define classes of viruses. Multi-scale approaches based on information theory[8] or repeated application of Newman's modularity[9] allow the study of genome similarity networks at multiple taxonomic levels. Local algorithms, such as OSLOM,[10] can detect overlapping modules (e.g. those resulting from mosaic genomes) and remove nodes whose module assignation is poorly supported statistically (e.g., single members of new taxa that occasionally share widespread genes with otherwise unrelated groups). Module detection in bipartite networks often involves Barber's modularity maximization[11] although relevant insights into the patterns of gene sharing and transmission can be obtained simply from the study of basic topological properties of the network.[3] The inference of viral groups from genome similarity networks might not differ much from unsupervised machine learning techniques but the more realistic representation of the virus world as a bipartite network of gene sharing makes network-based approaches more powerful at dealing with decaying degrees of genomic similarity at long evolutionary distances, as well as with poorly sampled taxa.

Genome similarity networks are more compact and easier to analyze than their bipartite counterparts but have several limitations. First, they lack information on the kind of genes that make 2 genomes similar, making it difficult to discriminate between cases of shared host-related genes and shared ancestral genes. Moreover, some properties of the final network may depend on the particular measure used to evaluate genome similarity. Finally, the projection of bipartite networks can lead to structural artifacts, such as spurious scale-free degree distributions.[12]

Despite these limitations, genome similarity networks have been successfully applied to bacteriophages to reveal the internal structure of this group of viruses[13] and to assign newly discovered phages to established families.[14] More recently, a large collection of viruses with dsDNA genomes has been studied under the framework of bipartite networks.[15] The analysis of the network showed that most dsDNA viruses belong to one of 2 major groups, each of which includes viruses from the 3 domains of life and is characterized by a distinct major capsid protein. Dissection of those groups leads to a hierarchy of subgroups which is consistent, despite some exceptions, with the established taxonomy of viruses. The hierarchical organization of the dsDNA virus world is founded on 3 classes of conserved genes: i) hallmark genes, such as capsid proteins, maturation proteases and packaging ATPases, that characterize and distinguish the 2 major viral groups, ii) connector genes, such as the baseplate proteins of myoviruses, that are shared by multiple subgroups within a group, and iii) signature genes that are highly specific to sets of related viruses (Fig. 2). Notably, most viruses that infect Archaea do not fall into the 2 major groups of dsDNA viruses and form a more fragmented network that is only weakly connected to the rest of the dsDNA virosphere, apparently reflecting the existence of stronger barriers to HGT among distinct groups of archaeal viruses and especially between viruses of archaea and bacteria.[16] In general, the pattern of connections is poorly conserved in more than 80% of the gene families of the bipartite virus network, in accord with the major role of HGT in virus evolution.

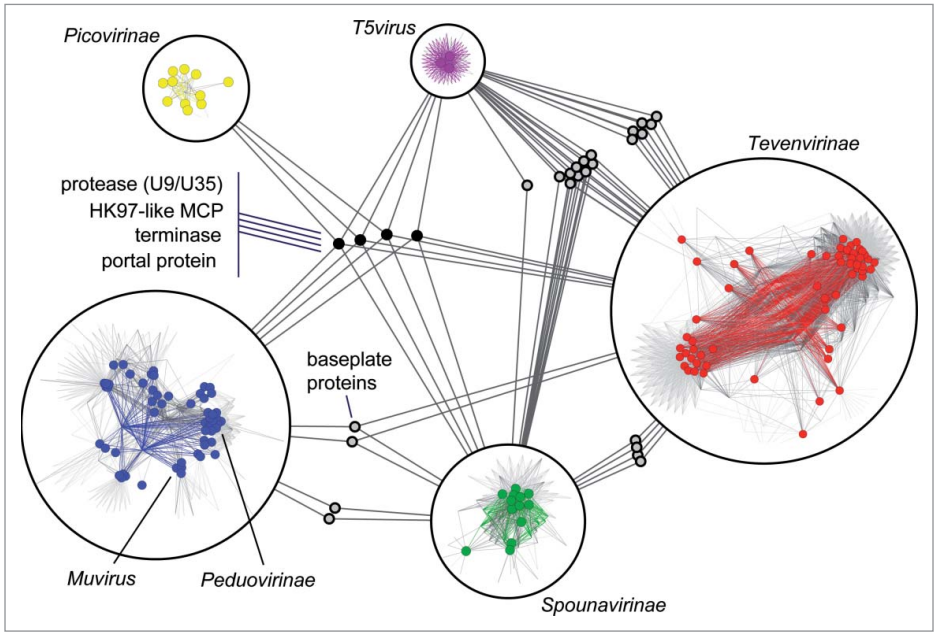


Figure 2. Hierarchical structure of a portion of the bipartite network for tailed bacteriophages (order *Caudovirales*). On the small scale, sets of related viruses and their associated gene families form densely connected modules. Within a module, genome nodes are represented as colored circles, whereas gene family nodes are denoted by the points where the edges (gray and colored lines) join. Colored edges connect the genomes of a module with the module's signature genes. On the large scale, modules connect with each other through shared connector genes, represented here as small gray circles. The 4 hallmark genes that are shared by most members of the order *Caudovirales* occupy a central position in the network (small black circles). This portion of the network corresponds to modules 9a, 9d, 12, 13, and 18 from ref. 15. MCP, major capsid protein.

As new tools for analysis of bipartite networks are developed, it soon will become possible to extend the network analyses to the entire virosphere and compare the findings from this approach with the patterns observed for viral hallmark genes.[17] From a complementary perspective, multilayer networks, such as the 2-layer representation of the dsDNA virosphere described above, have recently become a focus of network science.[18] Although technically challenging, a detailed analysis of the 2-layer network is a promising direction that will integrate sequence similarity and gene sharing in a unified framework. Eventually, additional layers accounting for host range, geographic location and environmental conditions would allow integration of information on genome evolution with ecological data, eventually resulting in a comprehensive picture of virus evolution.[19]

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

Funding

This work was funded by US Department of Health and Human Services (Intramural funds).

References

[1] Puigbò P, Wolf YI, Koonin EV. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. J Biol 2009; 8 (6):59; PMID:19594957; http://dx.doi.org/10.1186/jbiol159
[2] Koonin EV, Dolja VV. Virus world as an evolutionary network of viruses and capsidless selfish elements. Microbiol Mol Biol Rev 2014; 78(2):278-303; PMID:24847023; http://dx.doi.org/10.1128/MMBR.00049-13
[3] Corel E, Lopez P, Méheust R, Bapteste E. Network-thinking: Graphs to analyze microbial complexity and evolution. Trends Microbiol 2016; 24(3):224-37; PMID:26774999; http://dx.doi.org/10.1016/j.tim.2015.12.003
[4] Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 2000; 28 (1):33-6; PMID:10592175; http://dx.doi.org/10.1093/nar/28.1.33
[5] Li L, Stoeckert CJ, Jr., Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 2003; 13(9):2178-89; PMID:12952885; http://dx.doi.org/10.1101/gr.1224503
[6] Fortunato S. Community detection in graphs. Phys Rep 2010; 486(3–5):75-174; http://dx.doi.org/10.1016/j.physrep.2009.11.002
[7] Fortunato S, Hric D. Community detection in networks: A user guide. Phys Rep 2016; 659:1-44; http://dx.doi.org/10.1016/j.physrep.2016.09.002
[8] Rosvall M, Bergstrom CT. Multilevel Compression of Random Walks on Networks Reveals Hierarchical Orga-

nization in Large Integrated Systems. Plos One 2011; 6 (4):e18209; PMID:21494658; http://dx.doi.org/10.1371/journal.pone.0018209

[9] Zhang P, Moore C. Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. Proc Natl Acad Sci U S A 2014; 111(51):18144-9; PMID:25489096; http://dx.doi.org/10.1073/pnas.1409770111

[10] Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S. Finding statistically significant communities in networks. Plos One 2011; 6(4):e18961; PMID:21559480; http://dx.doi.org/10.1371/journal.pone.0018961

[11] Marquitti FMD, Guimaraes PR, Pires MM, Bittencourt LF. MODULAR: software for the autonomous computation of modularity in large network sets. Ecography 2014; 37(3):221-4; http://dx.doi.org/10.1111/j.1600-0587.2013.00506.x

[12] Montañez R, Medina MA, Solé RV, Rodríguez-Caso C. When metabolism meets topology: Reconciling metabolite and reaction networks. Bioessays 2010; 32(3):246-56; PMID:20127701; http://dx.doi.org/10.1002/bies.200900145

[13] Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Reticulate representation of evolutionary and functional relationships between phage genomes. Mol Biol Evol 2008; 25(4):762-77; PMID:18234706; http://dx.doi.org/10.1093/molbev/msn023

[14] Jang HB, Fagutao FF, Nho SW, Park SB, Cha IS, Yu JE, Lee JS, Im SP, Aoki T, Jung TS. Phylogenomic network and comparative genomics reveal a diverged member of the PhiKZ-related group, marine vibrio phage PhiJM-2012. J Virol 2013; 87(23):12866-78; PMID:24067958; http://dx.doi.org/10.1128/JVI.02656-13

[15] Iranzo J, Krupovic M, Koonin EV. The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. MBio 2016; 7(4):e00978-16; PMID:27486193; http://dx.doi.org/10.1128/mBio.00978-16

[16] Iranzo J, Koonin EV, Prangishvili D, Krupovic M. Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsidless mobile elements. J Virol 2016; 90(24):11043-55; PMID:27681128; http://dx.doi.org/10.1128/JVI.01622-16

[17] Koonin EV, Dolja VV, Krupovic M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. Virology 2015; 479–480:2-25; PMID:25771806; http://dx.doi.org/10.1016/j.virol.2015.02.039

[18] Kivela M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. Multilayer networks. J Comp Net 2014; 2(3):203-71; http://dx.doi.org/10.1093/comnet/cnu016

[19] Fondi M, Karkman A, Tamminen MV, Bosi E, Virta M, Fani R, Alm E, McInerney JO. "Every gene is everywhere but the environment selects": Global geolocalization of gene sharing in environmental samples through network analysis. Genome Biol Evol 2016; 8 (5):1388-400; PMID:27190206; http://dx.doi.org/10.1093/gbe/evw077