DGraph Clusters Flaviviruses and β-Coronaviruses According to Their Hosts, Disease Type, and Human Cell Receptors

Bioinformatics and Biology Insights
Volume 15: 1–9
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11779322211020316

Benjamin A Braun[1], Catherine H Schein[2,3] and Werner Braun[2,3]

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]Department of Biochemistry and Molecular Biology, The University of Texas Medical Branch, Galveston, TX, USA. [3]Institute for Human Infections and Immunity, The University of Texas Medical Branch, Galveston, TX, USA.

ABSTRACT

MOTIVATION: There is a need for rapid and easy-to-use, alignment-free methods to cluster large groups of protein sequence data. Commonly used phylogenetic trees based on alignments can be used to visualize only a limited number of protein sequences. DGraph, introduced here, is an application developed to generate 2-dimensional (2D) maps based on similarity scores for sequences. The program automatically calculates and graphically displays property distance (PD) scores based on physico-chemical property (PCP) similarities from an unaligned list of FASTA files. Such "PD-graphs" show the interrelatedness of the sequences, whereby clusters can reveal deeper connectivities.

RESULTS: Property distance graphs generated for flavivirus (FV), enterovirus (EV), and coronavirus (CoV) sequences from complete polyproteins or individual proteins are consistent with biological data on vector types, hosts, cellular receptors, and disease phenotypes. Property distance graphs separate the tick- from the mosquito-borne FV, cluster viruses that infect bats, camels, seabirds, and humans separately. The clusters correlate with disease phenotype. The PD method segregates the β-CoV spike proteins of severe acute respiratory syndrome (SARS), severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and Middle East respiratory syndrome (MERS) sequences from other human pathogenic CoV, with clustering consistent with cellular receptor usage. The graphs also suggest evolutionary relationships that may be difficult to determine with conventional bootstrapping methods that require postulating an ancestral sequence.

KEYWORDS: Alignment free clustering, Physical-chemical property (PCP) consensus, Property distances (PD) of viral sequences, Enterovirus classification, SARS origin

RECEIVED: October 2, 2020. **ACCEPTED:** May 5, 2021.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by grants from the National Institutes of Health (grant nos. AI137332 to W.B., AI064913 to W.B./C.H.S., and AI105985 to C.H.S.). The computational resources of the Sealy Center for Structural Biology and Molecular Biophysics were also used in this project.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Werner Braun, Department of Biochemistry and Molecular Biology, The University of Texas Medical Branch, Galveston, TX 77555, USA. Email: webraun@utmb.edu

Introduction

There are many methods being developed to handle and interpret the large amounts of sequence data available for viruses.[1-3] A well-done viral phylogeny is useful for suggesting the evolutionary relationships between viruses, their rates of change,[4] and may also alert one to tipping points where additional changes may result in significant phenotypic variation[5] or viral outbreak.[6,7] Determining the interrelatedness of virus sequences is perhaps most important for the design of broad spectrum vaccines and treatments for viral diseases.[8] Rooted trees based on such alignments imply a hierarchical, linear evolution and become difficult to interpret as the number of sequences increases. The limitations of these methods become obvious when determining relationships among the thousands of viral sequences now available. In practice, authors often resort to drawing 2D plots by hand to illustrate the interrelatedness of larger virus groups, although 2D-graphic, computational maps with BioLayout[9] or Cytoscape[10,11] are also used.

Here, we present a rapid graphical method for analyzing large data sets of related protein sequences that does not require prealignment or assumption of a common ancestor. DGraph can present conventional pairwise alignment scores, such as those from Clustal Omega,[12] or simple overall identity. However, the program's ability to generate "property distance" PD-graphs, based on physical-chemical properties (PCPs) of the amino acids[13] allows it to suggest more meaningful relationships among distantly related sequences. We have previously validated the PD method as a way to classify allergenic proteins and detect similar IgE epitopes.[14,15] We have shown that changes in the PCP values of key positions within flaviviral protein sequences correlate with significant phenotypic changes.[16,17] In addition to describing the details of the algorithms of the program, we show here its application to 3 diverse families of positive strand RNA viruses, flaviviruses (FVs),[3] enteroviruses (EVs),[18] and the β-coronaviruses (β-CoVs), which include severe acute respiratory syndrome (SARS), Middle East respiratory Syndrome (MERS), and the pandemic severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).[19] The results illustrate how PD-graphs of the viral sequences correlate with phenotype and suggest evolutionary relationships of distantly related viruses.

Material and Methods

The property distance (PD)

The peptide similarity search tool[14,15,20] was initially developed to find protein sequences in the Structural Database of Allergenic Proteins (SDAP)[21] containing user-specified peptide sequences. The search tool uses a novel technique to find similar sequences in the proteins by comparing the PCPs of the amino acids in the query and the target sequence. The differences in the PCP values in the 2 sequences are then measured by a PD. Briefly, 5 quantitative descriptors of PCPs are assigned to each of the 20 amino acids. The 5 descriptors E1 to E5 were derived by multidimensional scaling of 237 PCPs for the 20 naturally occurring amino acids, thus the main differences of all 237 properties for the 20 amino acids are reflected by the 5 descriptors E1 to E5.[13] These in turn represent groupings of PCPs such as hydrophobicity, size, or secondary structure propensities, charge, aromaticity, and size. The PD of 2 sequences A and B is then calculated as the average distance between the descriptor vectors E for corresponding amino acids, that is

$$PD(A,B)=\frac{1}{N}\sum_{i=1}^{N} d\left(\vec{E}(A_i),\vec{E}(B_i)\right)$$

where d computes the standard Euclidean or L2 distance and N is the length of sequence B, assumed to be the same as A in this equation (section "DGraph converts various sequence scoring functions or PD values into 2D maps" discusses the nonequal-length case.)

Identical sequences have a PD value of 0. Small PD values up to 4 typically indicate few substitutions with smaller values for conservative substitutions between the 2 sequences. Thus, the lower the PD value, the more closely related the sequences. In a database search of over 1500 allergenic protein sequences in SDAP, PD values of less than 8 are statistically significant for windows up to 12 amino acids and have been shown to correlate with immune recognition.[14] Additional statistical measures (z-scores) can be calculated to indicate the significance of a PD value comparing it to the distribution of PD values over all random matches using larger data sets.

The DGraph program

DGraph, as discussed below, can generate a 2D map based on any input value list. In default mode, if given a list of FASTA formatted sequences in a text file as input, it calculates the pairwise PD values for the sequences and graphically presents their similarities. The resulting "PD-graph" represents the sequences as nodes in a 2D map where the distances of the edges between nodes, "representative distances" are fitted to the PD values. Alternatively, if similarity scores are given as input, "sequence distances" are calculated from the similarity scores (see below in section "Algorithm to find the optimal configuration of the nodes") and the representative distances between nodes are fitted to these sequence distances. Whether fitting to internally calculated PD values, to sequence distances translated from similarity scores, or some other user-supplied distances, DGraph calculates a metric of the match of the representative distances between nodes in the generated graphic to the user-supplied distances and minimizes this metric.

Calculating PD values with a sliding window

For comparing sequences of amino acids that are not matched one-to-one in length, we instead define a metric comparing shorter fixed-length subsequences, or "windows." We define the windowed PD between 2 sequences to be the PD between the least distant pair of subsequences having some length wSize (for the window size). The windowed PD measures the sequence similarity of the most conserved portion of the 2 sequences of length wSize. Note that windowed PD is 0 between 2 sequences having an identical string of amino acids at a conserved region of length wSize. The occurrence of this is exponentially less likely as wSize increases; for the results in this article, we use a wSize of 22 amino acids. Windowed PD can be computed naively by considering each pair of wSize subsequences in turn and computing their PD. A more efficient approach exploiting the linearity of PD is to slide a window of length wSize along and compute PD incrementally, for each offset of the shorter sequence along the longer one, and keeping track of the smallest PD value found.

Algorithm to find the optimal configuration of the nodes

Similarity scores from alignment programs, such as Clustal OMEGA,[12] MUSCLE,[22] or T-Coffee[23] are translated into distances as

$$d_{\text{DGraph}}=\frac{1}{\left(\max\left\{SS,\sqrt{h}\right\}\right)^2}$$

where SS is a similarity score, and parameter h determines the minimum similarity cutoff. Similarity scores below h^0.5 are mapped to the maximum distance of $1/h$ in this equation, while larger similarity scores have an inverse square law with distance. The squaring reduces distances between highly similar sequences, which encourages visual clustering of similar sequences in the final figure DGraph produces.

Similarity scores SS are translated into sequence distances by the equation given above, and then these sequence distances are used to compute a measure, U, to fit optimally the representative distances between nodes to these distances

$$U=\sum_{(i,j)\varepsilon\Omega}\left(\frac{d\left(\overrightarrow{x_i},\overrightarrow{x_j}\right)-d_{i,j}}{d_{i,j}+1}\right)^2=\sum_{(i,j)\varepsilon\Omega}u_{i,j}^{\;2}$$

where the sum is over the set Omega of pairs of distinct sequences i, j, and the terms in the numerator are the representative distances and the sequence distances between those 2 sequences, respectively. U is a sum of the squared relative error of each pair of the sequence's representation placement versus the sequence distance, where the relative error ($u_{i,j}$) is computed with an adjusted denominator of $d_{i,j} + 1$ to account for small distances.

As described, optimizing U results in a figure determined primarily by the most distant pairs of sequences. To better represent relationships between closely related sequences, we optimize a distance-adjusted U^*, as follows: Define $u^*_{i,j}$, equal to $u_{i,j}$, but divided by $d(x_i,x_j)$, represent a distance only when the representation distance is at least 0.001% of the figure diameter, and $U^* = \mathrm{sum}(u^{*}_{i,j}{}^{2})$. In addition, we remove from Omega any pair (i,j) with a PD or user-defined score-based distance larger than a configurable maximum distance comparison cutoff (the default value is 14). Sequences with no distance below this cutoff to the rest of the figure then become "islands" that are removed before optimization begins. These steps are both intended to make the resulting DGraph figure more faithfully represent short distances.

Initially, DGraph creates randomly placed nodes for each sequence. DGraph then minimizes U^* using a gradient descent approach. That is, at each step, each node's position x_i is shifted by an amount proportional to $-u^*_{i,j}$ along each direction ($x_j - x_i$). To damp oscillations and promote convergence, we add a momentum vector p_i to each node, and apply the contribution of each $u^*_{i,j}$ to the momentum vectors rather than the position directly. Thus, at each step, the position of each node x_i and its associated momentum vector p_i are updated as follows

$$\overrightarrow{p_i} \leftarrow (1-f)\overrightarrow{p_i} + \frac{\Delta t}{m}\sum_{j\neq i}\left(-u^{*}_{i,j} \times \left(\overrightarrow{x_j} - \overrightarrow{x_i}\right)\right),$$
$$\overrightarrow{x_i} \leftarrow \overrightarrow{x_i} + \Delta t \times \overrightarrow{p_i}$$

where small positive coefficients time-step (delta t), mass (m), and friction (f) are parameters of the descent.

Use of the program and utility tools

Parameters of the optimization, such as time-step, mass, friction, and maximum distance comparison cutoff, can be user specified with default values of 0.001, 0.01, 0.008, and 14. DGraph can be run interactively to fine-tune these parameters and to see a live view of the optimization. DGraph can be run with a helper script that runs the optimization multiple times, potentially resulting in different figures. The script compares the runs by optimization score, and normalizes the orientation of the final locations of the nodes (by rotation and reflection) to produce a consensus graphic (this uses the Kabsch algorithm, for details see[24]). The user can customize how sequences are labeled on the generated graphic, such as including the whole of short sequences (when graphing peptides, see[25]) or FASTA names as labels. The user can also apply custom coloring to the graphic including a color gradient for the line segments between nodes based on PD value or user-defined distance, and files of per-node colors useful for annotating sequence properties such as phenotype.

Results

DGraph converts various sequence scoring functions or PD values into 2D maps

In previous studies, we established and validated the correlation of PD values for peptide segments with IgE binding affinities.[14] We therefore expect that PD-graphs would also be useful to predict antibody cross-reactivities of different viral species. In Supplementary Figure S1, we illustrate this feature for the PCP-consensus sequence 7P8 of the domain 3 of the E-protein of the 4 DENV viruses. The consensus domain 7P8 was designed as a potential vaccine candidate against the 4 major types of DENV and was recognized by antibodies generated against all 4. In the PD-graph 7P8 is located near the middle of all 4 viruses (Supplementary Figure S1).

An alignment of 49 different FV was then used to illustrate the flexibility of the program. Figure 1 shows the 2D-maps for 3 different sets of similarity data for viral sequences based on different metrics and lengths of proteins included in the calculation. The first column shows DGraph output calculated from previously calculated Clustal W alignment scores. The other 2 columns show 2 different ways to use PD values as a metric for generating the maps. The middle column shows the result of computing the PD value between each pair of sequences after a multiple alignment, which causes each sequence to be the same length by inserting alignment gaps. The last column shows results using instead a "sliding window" of 22 amino acids, by sliding every window in the shorter sequence along the longer sequence to find a best match, that is, the one with the lowest pairwise PD value and computing the average PD values of the matches.

The bottom frames of the figure show the results using a 60 amino acid region of the EV 3B-3C protein interface (see Supplemental Table S2), covering the viral protein linked to the genome (VPg)[26] and an area in 3C that contains a vestigial additional VPg-like sequence. For all 3 sets of viral sequences, the maps generated by the PD values show more distinct groupings than the maps generated by the ClustalW scores. They illustrate that the program can be used to rationally cluster even very long sequences (the whole viral genomes of the FV) as well as short sequences (the VPg area of the EVs).

The PD-graphs correlate with vector and host competence

The map from the top row (middle panel) in Figure 1 is further annotated and highlighted in more detail in Figure 2. The PD metrics shows a clear division of the tick from the mosquito-borne or no-known vector and Rio Bravo group viruses. The map also clusters viruses that infect bats, camels

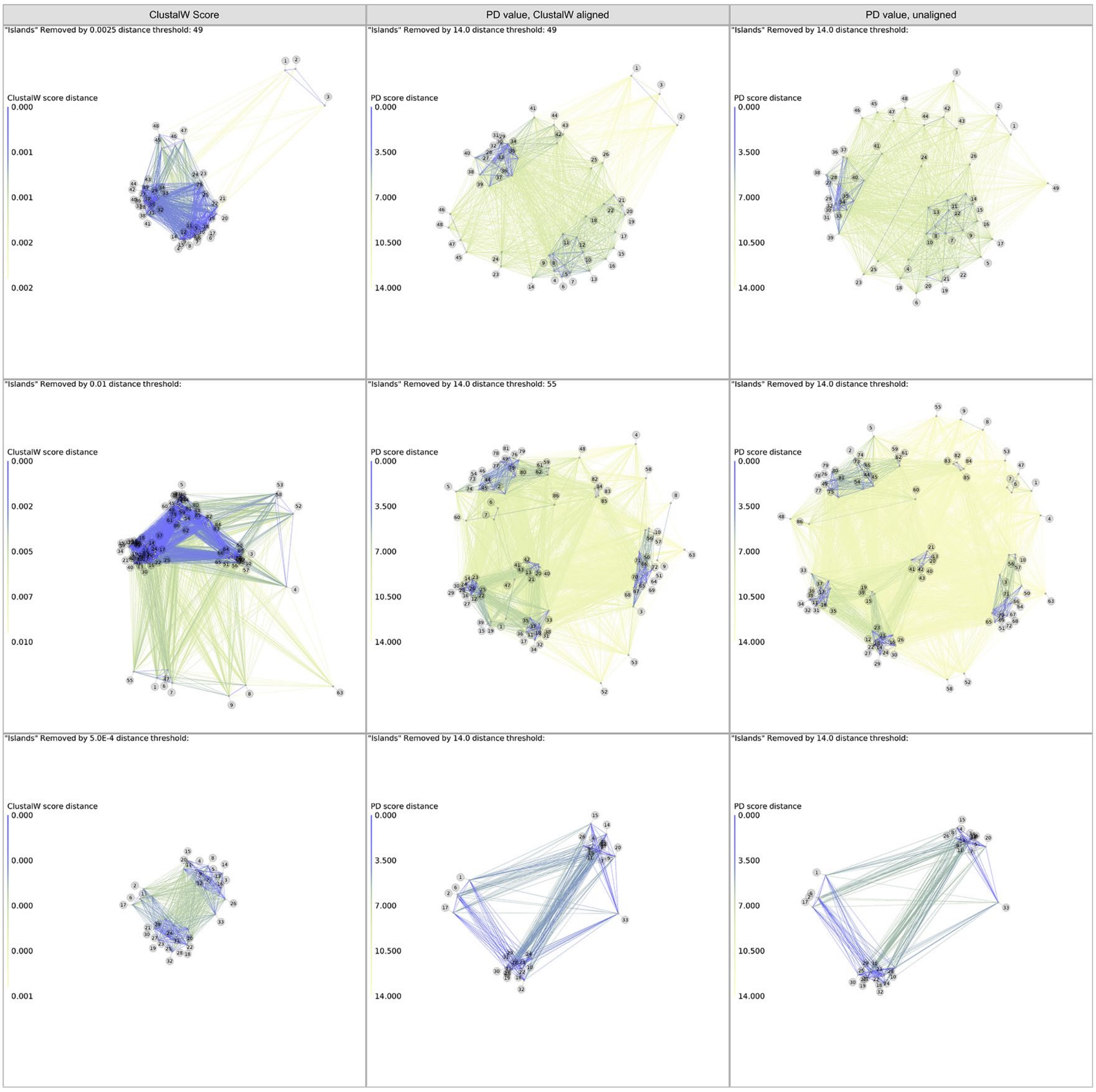
ClustalW Score
PD value, ClustalW aligned
PD value, unaligned
"Islands" Removed by 0.0025 distance threshold: 49
"Islands" Removed by 14.0 distance threshold: 49
"Islands" Removed by 14.0 distance threshold:
ClustalW score distance
0.000
0.001
0.001
0.002
0.002
PD score distance
0.000
3.500
7.000
10.500
14.000
"Islands" Removed by 0.01 distance threshold:
"Islands" Removed by 14.0 distance threshold: 55
"Islands" Removed by 14.0 distance threshold:
ClustalW score distance
0.000
0.002
0.005
0.007
0.010
"Islands" Removed by 5.0E-4 distance threshold:
"Islands" Removed by 14.0 distance threshold:
"Islands" Removed by 14.0 distance threshold:
ClustalW score distance
0.000
0.000
0.000
0.000
0.001

Figure 1. Top row: screen shots of the output of DGraph for the analysis of the 49 polyproteins of flaviviruses (Supplementary Table S1) with the ClustalW score (left panel), and the PD values of 22 residue windows with aligned (middle panel) and unaligned sequences (right panel). Middle row: screen shots from DGraph for the analysis of 86 sequences of the NS2a protein (all ~234 aa long) of various FV, based on ClustalW scores and their pairwise PD values. The sequences were automatically downloaded from a Blast search, identical sequences removed and the resulting FASTA files subjected to Clustal W analysis or our PD-based method with minimal involvement (except that the sequence headers were manually shortened and the sequences inspected to remove fragments). The resulting maps, which are similar to how protein family (PFAM) B families are generated, illustrate how PD-maps show a finer distinction among the viruses than simple Clustal scores using an unsupervised alignment. Bottom row: screen shots from DGraph for the clustering of enteroviruses. A 60 amino acid sequence around the VPg protein (22 amino acids) of human and a few animal enteroviruses was used as input. The 2D plots clearly separate the simian viruses from the human ones.
FV indicates flavivirus; PD, property distance.

(Kadam), and seabirds separately from those infecting humans, such as the Greek goat (node 29) and Turkish sheep encephalitis viruses (node 30) group. This observation also holds when the analysis is based on a smaller section of the virus, the 230 amino acid NS2a protein (middle row of Figure 1). The PD-graph is consistent with the patterns of insertions and deletions we previously noted within the E-protein that mark the FVs[16] according to species and disease specificity, as well as other phylogenic methods. With regard to the latter, hemorrhagic Yellow Fever (YF) virus (node 26) lies near the 4 DENV serotypes (nodes 19–22), which can cause hemorrhagic disease with fatal consequences in children. Interestingly, Zika virus (node 18) falls exactly between the mosquito-borne encephalitic and hemorrhagic groups, consistent with its cross-reactivity with DENV.[27] An identity matrix for the E-protein domain 3 region of Zika compared to WNV, DENV strains, and our 7P8 PCP-consensus protein that binds antibodies to all 4 DENV serotypes[28] illustrates how Zika indeed lies between the encephalitic and hemorrhagic FV, with >50% identity to all (Supplementary Figure 1). While Zika infections generally cause mild disease, they can also result in Guillain-Barré syndrome and

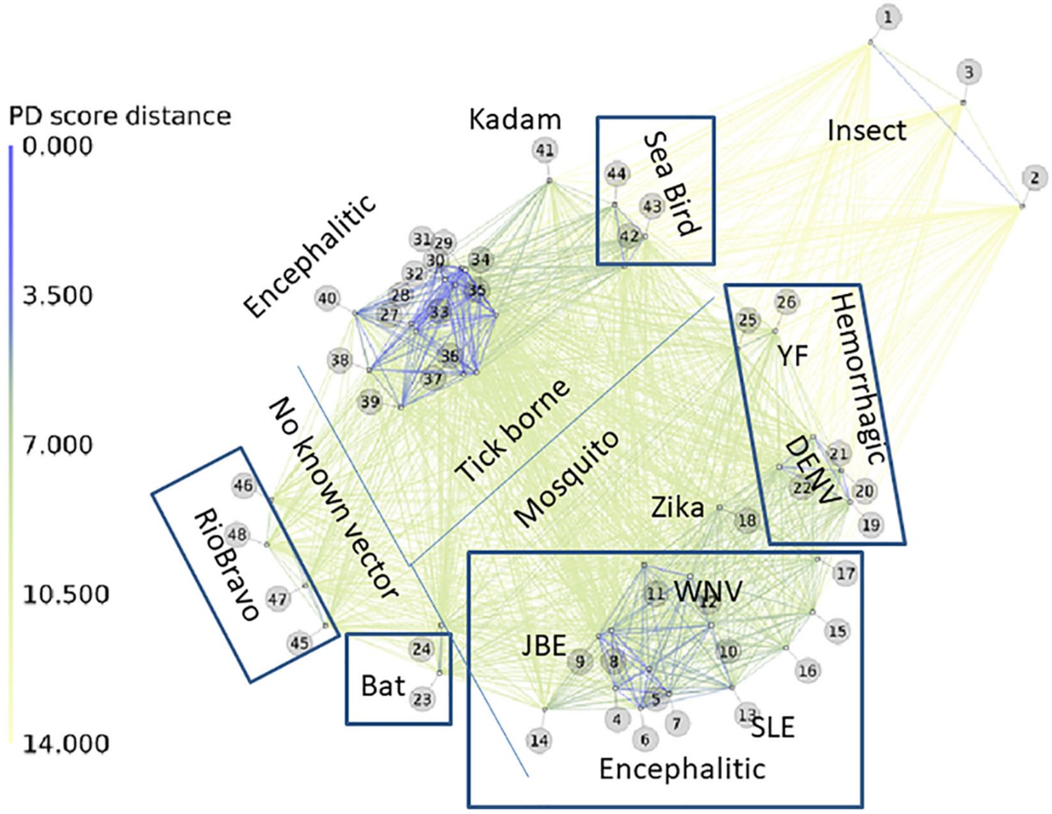

Figure 2. Annotated PD-graph of flavivirus phylogeny from the polyproteins of 49 viral species. Each number in the plot corresponds to the viruses listed in Table S1 with FlaviTrack ID, Genbank, species name, and length of amino acids. Starting from the unaligned FASTA sequences of the polyproteins, the program calculated pairwise PD values of all corresponding 22-residue segments for the polyproteins. The average PD values of the windows were used as a metric for the similarity of each pair. Lines in the figure indicate the degree of relatedness of the sequences, with a color code from blue to green as indicated with the PD values on the left of the figure. Blue thick lines indicate highly related flaviviruses. Divider lines and boxes were drawn by hand to emphasize the phenotypic groupings of the viral species.
PD indicates property distance.

microcephaly if contacted by a pregnant mother. Zika's nearest neighbor in the plot is the encephalitic virus, Rocio.[29]

The bottom part of Figure 1 and the annotated PD-graph (Figure 3) show that even short segments of EV sequences were sufficient to separate simian from human isolates. The EV are a nonenveloped group of +strand RNA viruses that includes poliovirus, coxsackie, and many other human pathogens, do not depend on arthropod vectors for transmission. The separation by the 2 parameters (Figure 1, bottom row), Clustal or PD, generated similar clusters of these viruses, whereby the PD-graph more clearly separated the 4 simian viruses that were distributed throughout the (unaligned) FASTA sequence list. DGraph of the Clustal values separated all the sequences into 2 large clusters, while there was more differentiation using PD values. The overall separation shown is consistent with genetic classifications done with other areas of the sequences[30] and references included in Supplementary Material). Although classifying EVs according to disease type has proved very difficult[31] and will not be attempted here, the clustering cleanly separates the C-type EV, which include polioviruses and many coxsackie A strains, from the B-type, which includes the Coxsackie virus-B, and other EV (Figure 3 and Supplementary Figure S2). This distinction is important, as the B-type strains are associated with many severe illnesses and cardiomyopathy. The identification of CVA 22 as an outlier that has close association with C-type EV, but also linkage to the B-type, is also consistent with reported recombination events.

PD-graph accurately separate β-CoV according to disease type and receptor used

Severe acute respiratory syndrome coronavirus 2 is known to be closely related, in its sequence, structure,[32] receptor binding[33] and epitopes recognized by neutralizing antibodies isolated from survivors[34-36] to the SARS-CoV-1 virus that caused many deaths in a brief epidemic that ended in Asia in 2003.[37] It is more distantly related to the lethal MERS.[38]

As an additional test of the program, 314 sequences of the spike protein for diverse β-CoV were downloaded from the ViPR database, and a single unaligned FASTA file used as input to the program. As Figure 4 shows, the resulting PD-graph, using the 22 amino acid window method, separates the β-CoV into distinct clusters according to their cellular receptor. The SARS virus from the 2002-2003 epidemic forms

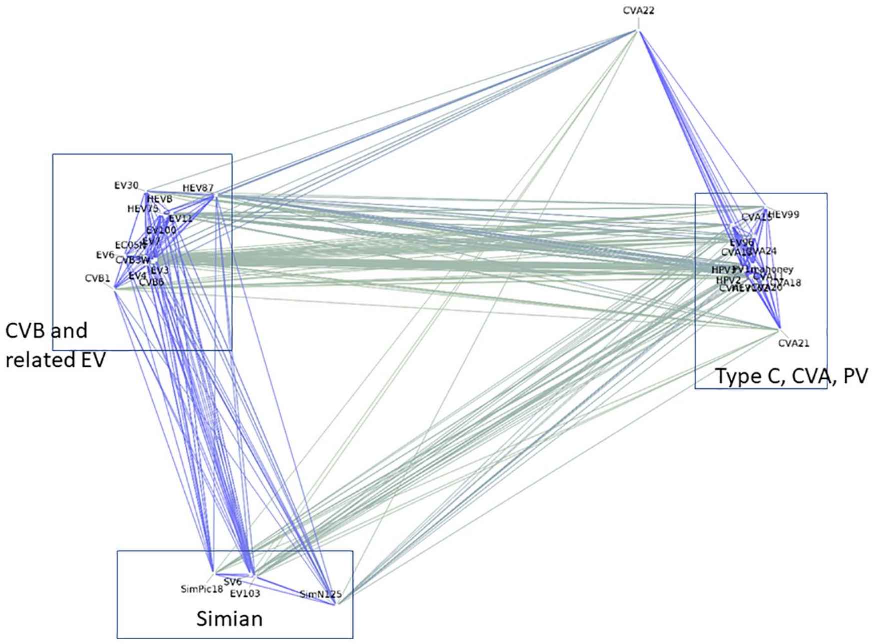

Figure 3. Annotated 2D-map for selected enterorviruses based only on a short region encompassing their VPg sequences and part of the 3C protein (See Supplementary Table S2 for sequence details and what is known about disease phenotypes). The clustering clearly separates the simian and baboon from the human viruses as well as the PV and CVA viruses ("EV type C").
EV indicates enterovirus.

a tight cluster with the SARS-CoV-2 from the current pandemic. This cluster has very high PD values to any of the other β-CoV clusters, including the MERS viruses of the 2012-2013 outbreak, which use a different cellular receptor, the DPP4 protein and groups with many camel isolates. These in turn are cleanly separated from the cluster of strains related to the less lethal CoV OC43 (which uses MHC-class 1 molecules as a fusion receptor[39]) and bat strains related to HKU4 and HKU5. The central nodes of the graph (red arrows) are viruses from bats that use the same DPP4 receptor as MERS.[40]

The viruses closest to SARS-CoV-2 are human SARS sequenced in 2003, M15, an isolate from a fatal human SARS infection that was passaged in mice and sequenced in 2003 at the NIH, and a bat virus isolated in 2007 in China. The low PD of these viruses to one another, as well as their very high PD to any of the other clusters, suggests that the more recent virus is an evolved form of that circulating in 2002-2003. As the annotations show, both use the human ACE2 receptor for cell binding, in contrast to the receptors identified for other β-CoV.

Discussion

Most phylogenetic analysis of viral sequences starts with multiple sequence alignments. Aligning very diverse sequences, especially those containing multiple repeats or insertions, can be quite difficult.[41] DGraph, as a flexible sequence analysis tool for protein sequences, provides a valuable first step to obtain an overview of related viral species without the need for an alignment. DGraph is unique in that it implements the PCP descriptors and PD calculations which were previously validated in our work in comparing allergenic proteins and their epitopes.[42,43] No hypothesis about an ancestral sequence is required to follow interstrain differentiation. As Figure 1 illustrates, simply starting with a list of unaligned sequences in FASTA format, one can rapidly determine the interrelatedness of sequence data from 2 different virus families, the FV and EV, whereby the PD approach can give more specific clustering than simple clustal scores. As Figures 2 to 4 show, automatic calculation of the PD values between even large groups of viral sequences yields PD-graphs consistent with what is known of the their vector, host range, receptor type (particularly for the β-CoV, Figure 4) and disease phenotypes.

The approach in DGraph for the placement of the nodes in a 2D map follows previously suggested force-directed methods.[44,45] The computational graph drawing also resembles other commonly used approaches for visualization of network connectivity, such as BioLayout,[9] or Cytoscape.[10,11] Other approaches such as DIALIGN2 (43) explore the importance of local alignments in avoiding the problems of global alignments for a diverse set of protein sequences.

Other methods for high-throughput approaches have been designed for large-scale analysis of protein sequences from genomic data, such as the derivation of clusters of orthologous groups (COGs) of protein sequences[46] and the Tribe-MCL method.[47] Both methods base their clustering on exhaustive pairwise sequence comparisons and define clusters as consistent sets of connected nodes. The automated Tribe-MCL method finds sequence clusters by a Monte Carlo approach from pairwise similarity scores and simulates random walks between the nodes with transition probabilities derived from the scores. Our DGraph approach solves the 2D-embedding

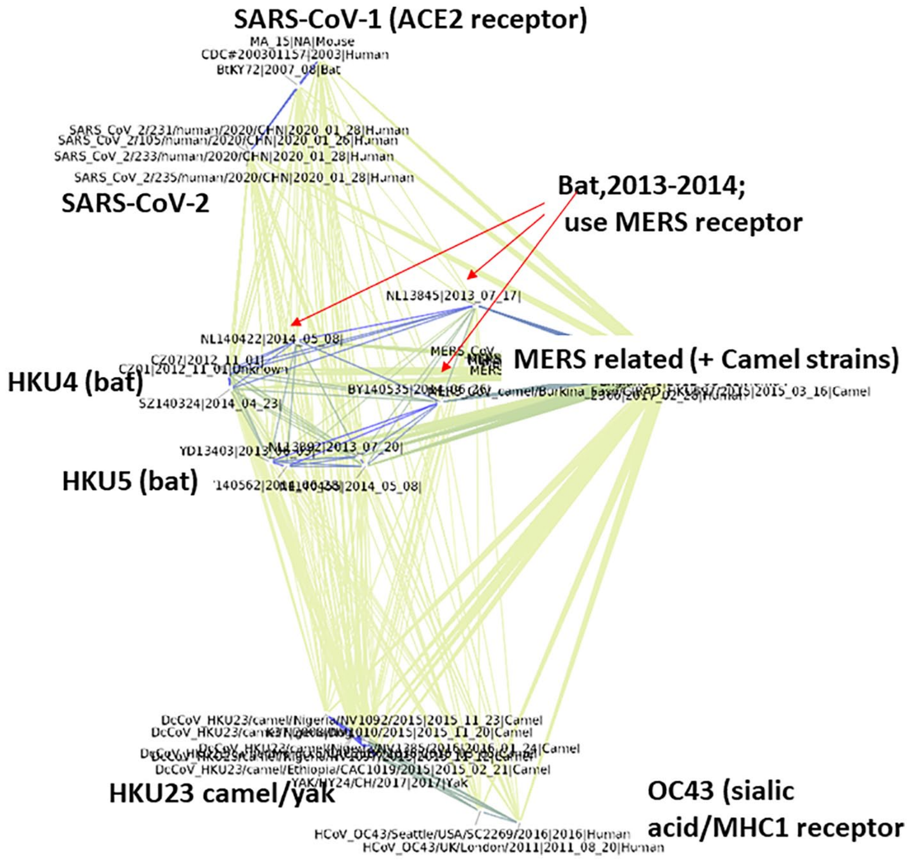
SARS-CoV-1 (ACE2 receptor)
MA_15|NA|Mouse
CDC#200301157|2003|Human
BtKY72|2007_08|Bat
SARS_CoV_2/231/human/2020/CHN|2020_01_28|Human
SARS_CoV_2/105/human/2020/CHN|2020_01_26|Human
SARS_CoV_2/233/human/2020/CHN|2020_01_28|Human
SARS_CoV_2/235/human/2020/CHN|2020_01_28|Human
SARS-CoV-2
Bat,2013-2014;
use MERS receptor
NL13845|2013_07_17|
NL140422|2014_05_08|
CZ07|2012_11_01|
CZ01|2012_11_01|Unknown
HKU4 (bat)
MERS_CoV
MERS
MERS related (+ Camel strains)
BY140535|2014_05_26|
camel/Burkina_Faso
2015_03_16|Camel
SZ140324|2014_04_23|
YD13403|2013
NL13892|2013_07_20|
HKU5 (bat)
140562|2014
2014_05_08|
DcCoV_HKU23/camel/Nigeria/NV1092/2015|2015_11_23|Camel
2010/2015|2015_11_20|Camel
DcCoV_HKU23/camel/Nigeria/NV1385/2016|2016_01_24|Camel
DcCoV_HKU23/camel/Ethiopia/CAC1019/2015|2015_02_21|Camel
YAK/HY24/CH/2017|2017|Yak
HKU23 camel/yak
OC43 (sialic
acid/MHC1 receptor
HCoV_OC43/Seattle/USA/SC2269/2016|2016|Human
HCoV_OC43/UK/London/2011|2011_08_20|Human

Figure 4. PD-graph groups SARS-CoV-1 and -2 spike proteins and distinguishes them from other circulating β-CoV strains. The 2020 isolates of SARS-CoV-2, which like the SARS viruses use the human ACE2 receptor, are closest to human SARS 2003, MA-15, from a human case in 2003 passaged in mice and a 2007 strain from a bat than they are to other circulating β-CoV strains. Annotations indicate grouping according to receptor type, where known. Red arrows show nodes, bat viruses BY140535 (HKU5 related), NL13845, and NL140422 strains that use the MERS receptor, DPP4, all sequenced in China in 2013/2014. MERS and MERS-related strains are highly similar, thus the labels of all these sequences overlap in the PD-graph which is annotated as MERS-related + camel strains. Blue lines show PD < 7 (low PD = more similar), other lines are PD < 14, whereby the thickness indicates the degree of relatedness.
CoV indicates coronavirus; MERS, Middle East respiratory syndrome; PD, property distance.

problem of distances by a force-directed approach. Combining different clustering options and visualization in one practical software package is a unique feature of DGraph which makes DGraph a useful exploratory tool for generating functional and evolutionary hypotheses.

The program's default mode can generate PD-graphs even of large numbers of unaligned sequences (such as the >300 used for Figure 4). While other methods can graphically present protein sequences in an alignment-free manner using numerical descriptors for the amino acids, and display them as connecting vectors in a curve in a 2D space,[48,49] they are most useful for comparing a few sequences. The DGraph program works from a list of unaligned sequences and can also use additional data, while allowing the user to adjust the program parameters to obtain results even for very distantly related sequences.

The major features of the 2D map by DGraph for FVs (Figure 2) are consistent with previously published phylogenetic relations between FVs. This includes our previous work,[50] where the major FV reference sequences were grouped using principal components analysis based on the sum of pairwise BLOSUM scores for the eigenvector decomposition (Figure 2 of Misra & Schein).[50] It is also consistent with a phylogenetic tree analysis of FVs using a Markov chain Monte Carlo analysis implemented in MrBayes[51] (Figure 1). Both studies agree on the 4 major group of FVs, the insect-specific, the tick-borne, the mosquito-borne, and no known-vector FVs and that there is a clear distinction between the encephalitic and hemorrhagic FV. These distinct groups are consistent with the results of DGraph (Figure 2). However, the 2D map of DGraph also can suggest potential evolutionary paths as described below or identify a central role for individual viruses as for example for YFV, which has connectivities to both the tick-borne and mosquito-borne viruses. Those relations are difficult to discern in a hierarchical representation of standard phylogenetical trees.

We see similar differentiation in both the EV (Figure 3). Here, the strains were separated cleanly into those of simian origin from the human strains. These in turn were separated into EV-Type-C and Type-B strains, which also carry some phenotypic information. For example, 2 of the strains that are identified as causing acute flaccid paralysis (Supplementary Table S2) are outliers within their type groups. However, the clusters also include strains that had no obvious effect in the people from whom they were isolated.

Property distance graphs can suggest evolutionary paths between distantly related viruses. PD-graphs emphasize the nonlinearity of viral evolution (ie, the seemingly random alteration pathways that lead to the many different strains of a given virus that occur over time). These pathways may be missed when using phylogenetic trees to model the evolution of viral groups, whereby there is no implied directionality in the connecting lines, which only represent the mathematical relationship between pairs. However, the PD-graph analysis of the FV (Figure 2) has features that correlate with suggested paths for the divergence of the mammalian pathogens. The clustering emphasizes the central position of YFV, relative to both mosquito- and tick-born viruses and the "mosquito only" viruses. This implies that the ability to circulate within their arthropod vectors may have taken priority during evolution of the mammalian pathogens. This may account for the relative stability of the YFV genome compared to that of the DENV types, which must have evolved under pressure from mammalian immune responses.[4] Another interesting feature of the PD-graph is the grouping according to known host even when isolated from geographically very distant places. Entebbe (node 23) and Yokose (node 24) viruses, which were isolated from bats in Uganda (ENTV)[52] and Japan (YOKV) group together, but also have strong connectivity to other mosquito-borne viruses. The YOKV sequence cross-reacts with antibodies in sera from humans infected with DENV or after vaccination with YFV[5] and, depending on the protein area chosen, is similar to many different mosquito-borne FV. Some of the early reports were not conclusive about the cross-reactivity of ENTV with other FV.[53,54] Property distance-graphs could provide testable hypotheses on FV cross-recognition, by comparing graphs made using inter-strain enzyme-linked immunosorbant assay (ELISA) values and PD values.

YFV. in turn has connectivity to both the tick- and mosquito-borne viruses, which subdivide into well-separated clusters according to their disease phenotype. The close relationship of the tick-born viruses to one another suggests that their ancestry is relatively recent or that other factors in the tick life cycle may constrain their evolution rate.[55,56] The distinct properties of the tick- versus the mosquito-borne viruses[57,58] illustrated by their neatly defined clusters, reflects these influences.

For β-COVs the evolutionary path is still debated and the zoonotic origin of the SARS-CoV-2 pandemic is still being investigated. However, most observations indicate that CoVs can move from one species to another, as for example from camels to man (for MERS) and from human to mink and vice versa for SARS-CoV-2[59], emphasizing the need for antigen based testing for animals that may be asymptomatic carriers[60] Our β-CoV PD-Graph (Figure 4) is consistent with those observations, separating the diverse strains mainly according to their receptor types and not to host types.

Conclusion

The DGraph program can be used to plot the interrelatedness of sequences according to PCP similarity and suggest evolutionary relationships, without needing an alignment or assuming a common ancestor. While the figures shown here illustrate the application to viral proteins, any group of sequences or numerical relationships of objects, including immunological metrics, can in principle be used as input to the program. We thus anticipate that it will find numerous uses for the increasing numbers of virus sequences, as well as those for many other areas and are herewith releasing a downloadable version of the program.

Author Contributions

BAB wrote the DGraph program and associated descriptive material, with consultation of WB, who derived the original PCP-based analysis method for protein sequences. CHS assembled the virus sequences, generated the PD-graphs of Figs. 2-4 and correlated the clusters with related published data. All 3 authors prepared and edited the paper.

Availability and Implementation

DGraph is written in Java, compatible with the Java 5 runtime or newer. Source code and executable is available from the GitHub website (https://github.com/bjmnbraun/DGraph/releases). Documentation for installation and use of the software is available from the Readme.md file at (https://github.com/bjmnbraun/DGraph).

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Suchard MA, Rambaut A. Many-core algorithms for statistical phylogenetics. *Bioinformatics*. 2009;25:1370-1376. doi:10.1093/bioinformatics/btp244.
2. Holmes EC, Grenfell BT. Discovering the phylodynamics of RNA viruses. *PLoS Comput Biol*. 2009;5:e1000505. doi:10.1371/journal.pcbi.1000505.
3. Simmonds P, Becher P, Bukh J, et al. ICTV virus taxonomy profile: Flaviviridae. *J Gen Virol*. 2017;98:2-3. doi:10.1099/jgv.0.000672.
4. Sall AA, Faye O, Diallo M, Firth C, Kitchen A, Holmes EC. Yellow fever virus exhibits slower evolutionary dynamics than dengue virus. *J Virol*. 2010;84:765-772. doi:10.1128/JVI.01738-09.
5. Tajima S, Takasaki T, Matsuno S, Nakayama M, Kurane I. Genetic characterization of Yokose virus, a flavivirus isolated from the bat in Japan. *Virology*. 2005;332:38-44. doi:10.1016/j.virol.2004.06.052.
6. Nelson MI, Edelman L, Spiro DJ, et al. Molecular epidemiology of A/H3N2 and A/H1N1 influenza virus during a single epidemic season in the United States. *Plos Pathog*. 2008;4:e1000133. doi:10.1371/journal.ppat.1000133.
7. Murcia PR, Baillie GJ, Daly J, et al. Intra- and interhost evolutionary dynamics of equine influenza virus. *J Virol*. 2010;84:6943-6954. doi:10.1128/JVI.00112-10.
8. Baker WS, Negi S, Braun W, Schein CH. Producing physicochemical property consensus alphavirus protein antigens for broad spectrum vaccine design. *Antiviral Res*. 2020;182:104905. doi:10.1016/j.antiviral.2020.104905.
9. Goldovsky L, Cases I, Enright AJ, Ouzounis CA. BioLayout(Java): versatile network visualisation of structural and functional relationships. *Appl Bioinformatics*. 2005;4:71-74. doi:10.2165/00822942-200504010-00009.
10. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13:2498-2504. doi:10.1101/gr.1239303.
11. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*. 2011;27:431-432. doi:10.1093/bioinformatics/btq675.

12. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci*. 2018;27:135-145. doi:10.1002/pro.3290.
13. Venkatarajan MS, Braun W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. *J Mol Mod*. 2001;7:445-453.
14. Ivanciuc O, Midoro-Horiuti T, Schein CH, et al. The property distance index PD predicts peptides that cross-react with IgE antibodies. *Mol Immunol*. 2009;46:873-883. doi:10.1016/j.molimm.2008.09.004.
15. Schein CH, Ivanciuc O, Braun W. Bioinformatics approaches to classifying allergens and predicting cross-reactivity. *Immunol Allergy Clin North Am*. 2007;27:1-27. doi:10.1016/j.iac.2006.11.005.
16. Danecek P, Lu W, Schein CH. PCP consensus sequences of flaviviruses: correlating variance with vector competence and disease phenotype. *J Mol Biol*. 2010;396:550-563. doi:10.1016/j.jmb.2009.11.070.
17. Danecek P, Schein CH. Flavitrack analysis of the structure and function of West Nile non-structural proteins. *Int J Bioinform Res Appl*. 2010;6:134-146. doi:10.1504/IJBRA.2010.032117.
18. Oberste MS, Maher K, Kilpatrick DR, Pallansch MA. Molecular evolution of the human enteroviruses: correlation of serotype with VP1 sequence and application to picornavirus classification. *J Virol*. 1999;73:1941-1948. doi:10.1128/JVI.73.3.1941-1948.1999.
19. Boni MF, Lemey P, Jiang X, et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat Microbiol*. 2020;5:1408-1417. doi:10.1038/s41564-020-0771-4.
20. Ivanciuc O, Braun W. Robust quantitative modeling of peptide binding affinities for MHC molecules using physical-chemical descriptors. *Protein Pept Lett*. 2007;14:903-916. doi:10.2174/092986607782110257.
21. Ivanciuc O, Schein CH, Braun W. SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res*. 2003;31:359-362. doi:10.1093/nar/gkg010.
22. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 2004;5:113. doi:10.1186/1471-2105-5-113.
23. Di Tommaso P, Moretti S, Xenarios I, et al. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res*. 2011;39:W13-W17. doi:10.1093/nar/gkr245.
24. Kabsch W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr*. 1976; A32:922.
25. Nesbit JB, Schein CH, Braun BA, et al. Epitopes with similar physicochemical properties contribute to cross reactivity between peanut and tree nuts. *Mol Immunol*. 2020;122:223-231. doi:10.1016/j.molimm.2020.03.017.
26. Schein CH, Ye M, Paul AV, et al. Sequence specificity for uridylylation of the viral peptide linked to the genome (VPg) of enteroviruses. *Virology*. 2015;484:80-85. doi:10.1016/j.virol.2015.05.016.
27. Barba-Spaeth G, Dejnirattisai W, Rouvinski A, et al. Structural basis of potent Zika-dengue virus antibody cross-neutralization. *Nature*. 2016;536:48-53. doi:10.1038/nature18938.
28. Bowen DM, Lewis JA, Lu W, Schein CH. Simplifying complex sequence information: a PCP-consensus protein binds antibodies against all four Dengue serotypes. *Vaccine*. 2012;30:6081-6087. doi:10.1016/j.vaccine.2012.07.042.
29. Medeiros DBA, Nunes MRT, Vasconcelos PFC, Chang GJ, Kuno G. Complete genome characterization of Rocio virus (Flavivirus: Flaviviridae), a Brazilian flavivirus isolated from a fatal case of encephalitis during an epidemic in Sao Paulo state. *J Gen Virol*. 2007;88:2237-2246. doi:10.1099/vir.0.82883-0.
30. Oberste MS, Maher K, Pallansch MA. Molecular phylogeny of all human enterovirus serotypes based on comparison of sequences at the 5’ end of the region encoding VP2. *Virus Res*. 1998;58:35-43. doi:10.1016/s0168-1702(98)00101-4.
31. Rosen L. Subclassification of picornaviruses. *Bacteriol Rev*. 1965;29:173-184.
32. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*. 2020;183:1735. doi:10.1016/j.cell.2020.11.032.
33. Walls AC, Xiong X, Park YJ, et al. Unexpected receptor functional mimicry elucidates activation of coronavirus fusion. *Cell*. 2019;176:1026-1039.e15. doi:10.1016/j.cell.2018.12.028.
34. Wec AZ, Wrapp D, Herbert AS, et al. Broad sarbecovirus neutralizing antibodies define a key site of vulnerability on the SARS-CoV-2 spike protein. *bioRxiv*. 2020. doi:10.1101/2020.05.15.096511.
35. Barnes CO, West AP Jr, Huey-Tubman KE, et al. Structures of human antibodies bound to SARS-CoV-2 spike reveal common epitopes and recurrent features of antibodies. *Cell*. 2020;182:828-842.e16. doi:10.1016/j.cell.2020.06.025.
36. Pinto D, Park YJ, Beltramello M, et al. Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature*. 2020;583:290-295. doi:10.1038/s41586-020-2349-y.
37. Rota PA, Oberste MS, Monroe SS, et al. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science*. 2003;300:1394-1399. doi:10.1126/science.1085952.
38. van Boheemen S, de Graaf M, Lauber C, et al. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *mBio*. 2012;3:e00473-12. doi:10.1128/mBio.00473-12.
39. Owczarek K, Szczepanski A, Milewska A, et al. Early events during human coronavirus OC43 entry to the cell. *Sci Rep*. 2018;8:7124. doi:10.1038/s41598-018-25640-0.
40. Luo CM, Wang N, Yang XL, et al. Discovery of novel bat coronaviruses in South China that use the same receptor as Middle East Respiratory Syndrome coronavirus. *J Virol*. 2018;92:e00116-18. doi:10.1128/JVI.00116-18.
41. Schein CH. Polyglutamine repeats in viruses. *Mol Neurobiol*. 2019;56:3664-3675. doi:10.1007/s12035-018-1269-4.
42. Lu W, Negi SS, Schein CH, Maleki SJ, Hurlburt BK, Braun W. Distinguishing allergens from non-allergenic homologues using physical-chemical property (PCP) motifs. *Mol Immunol*. 2018;99:1-8. doi:10.1016/j.molimm.2018.03.022.
43. Ivanciuc O, Garcia T, Torres M, Schein CH, Braun W. Characteristic motifs for families of allergenic proteins. *Mol Immunol*. 2009;46:559-568. doi:10.1016/j.molimm.2008.07.034.
44. Fruchterman T, Reingold E. Graph drawing by force-directed placement. *Softw-Pract Exp*. 1991;21:1129-1164.
45. Kobourov SG. Spring embedders and force directed graph drawing algorithms. *arXiv*. 2012;12013011, https://arxiv.org/pdf/1201.3011.pdf.
46. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science*. 1997;278:631-637. doi:10.1126/science.278.5338.631.
47. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*. 2002;30:1575-1584. doi:10.1093/nar/30.7.1575.
48. Randic M, Zupan J, Balaban AT, Vikic-Topic D, Plavsic D. Graphical representation of proteins. *Chem Rev*. 2011;111:790-862. doi:10.1021/cr800198j.
49. Hu H, Li Z, Dong H, Zhou T. Graphical representation and similarity analysis of protein sequences based on fractal interpolation. *IEEE/ACM Trans Comput Biol Bioinform*. 2017;14:182-192. doi:10.1109/TCBB.2015.2511731.
50. Misra M, Schein CH. Flavitrack: an annotated database of flavivirus sequences. *Bioinformatics*. 2007;23:2645-2647. doi:10.1093/bioinformatics/btm383.
51. Moureau G, Cook S, Lemey P, et al. New insights into flavivirus evolution, taxonomy and biogeographic history, extended by analysis of canonical and alternative coding sequences. *PLoS ONE*. 2015;10:e0117849. doi:10.1371/journal.pone.0117849.
52. Kading RC, Kityo R, Nakayiki T, et al. Detection of Entebbe bat virus after 54 years. *Am J Trop Med Hyg*. 2015;93:475-477. doi:10.4269/ajtmh.15-0065.
53. Peat A, Bell TM. Entebbe bat salivary gland virus: electron microscopic study of morphology and development in new born mice. *Arch Gesamte Virusforsch*. 1970;31:230-236. doi:10.1007/BF01253757.
54. Kuno G, Chang GJ. Characterization of Sepik and Entebbe bat viruses closely related to yellow fever virus. *Am J Trop Med Hyg*. 2006;75:1165-1170.
55. Grard G, Moureau G, Charrel RN, et al. Genetic characterization of tick-borne flaviviruses: new insights into evolution, pathogenetic determinants and taxonomy. *Virology*. 2007;361:80-92. doi:10.1016/j.virol.2006.09.015.
56. Ruzek D, Gritsun TS, Forrester NL, et al. Mutations in the NS2B and NS3 genes affect mouse neuroinvasiveness of a Western European field strain of tick-borne encephalitis virus. *Virology*. 2008;374:249-255. doi:10.1016/j.virol.2008.01.010.
57. Lawrie CH, Uzcategui NY, Armesto M, Bell-Sakyi L, Gould EA. Susceptibility of mosquito and tick cell lines to infection with various flaviviruses. *Med Vet Entomol*. 2004;18:268-274. doi:10.1111/j.0269-283X.2004.00505.x.
58. Kofler RM, Hoenninger VM, Thurner C, Mandl CW. Functional analysis of the tick-borne encephalitis virus cyclization elements indicates major differences between mosquito-borne and tick-borne flaviviruses. *J Virol*. 2006;80:4099-4113. doi:10.1128/JVI.80.8.4099-4113.2006.
59. Oude Munnink BB, Sikkema RS, Nieuwenhuijse DF, et al. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science*. 2021;371:172-177. doi:10.1126/science.abe5901.
60. Schein CH, Levine CB, McLellan SCF, Negi SS, Braun W, Dreskin SC, Anaya SC, Schmidt J. Synthetic proteins for COVID-19 diagnostics Peptides 2021 (in press).