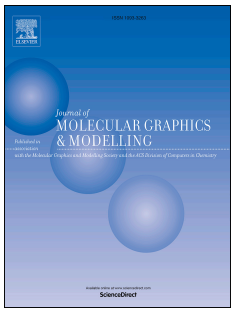
Accepted Manuscript



Visualization of protein sequence space with force-directed graphs, and their application to the choice of target-template pairs for homology modelling

Dylan J.T. Mead, Simón Lunagomez, Derek Gatherer

PII: S1093-3263(19)30333-X

DOI: https://doi.org/10.1016/j.jmgm.2019.07.014

Reference: JMG 7417

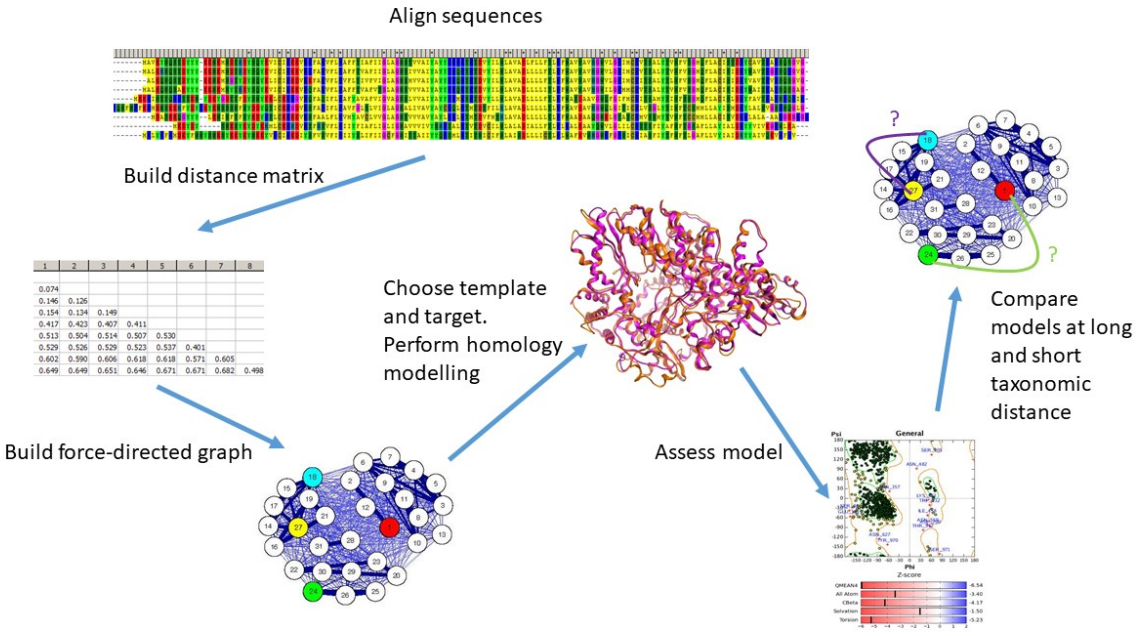
To appear in: *Journal of Molecular Graphics and Modelling*

Received Date: 7 May 2019

Revised Date: 23 July 2019

Accepted Date: 25 July 2019

Please cite this article as: D.J.T. Mead, Simó. Lunagomez, D. Gatherer, Visualization of protein sequence space with force-directed graphs, and their application to the choice of target-template pairs for homology modelling, *Journal of Molecular Graphics and Modelling* (2019), doi: https://doi.org/10.1016/j.jmgm.2019.07.014.



TITLE PAGE

Visualization of protein sequence space with force-directed graphs, and their application to the choice of target-template pairs for homology modelling.

Dylan J.T. Mead [1], *Simón Lunagomez* [2] *& Derek Gatherer* [1]* (surnames underlined)

[1] Division of Biomedical & Life Sciences, Faculty of Health & Medicine, Lancaster University, Lancaster LA1 4YT, UK.

[2] Department of Mathematics & Statistics, Lancaster University, Lancaster LA1 4YF, UK.

* Corresponding author: d.gatherer@lancaster.ac.uk, +44 1524 592900, Twitter: @DerekGatherer

Other author emails:

DJTM: dylanmead.dm@googlemail.com

SL: s.lunagomez@lancaster.ac.uk

Word count: 4652

Running Title: **Force-directed graphs, homology modelling and the structure–sequence gap**

Visualization of protein sequence space with force-directed graphs, and their application to choice of target-template pairs for homology modelling.

Dylan J.T. Mead [1], *Simón Lunagomez* [2] *& Derek Gatherer* [1]* (surnames underlined)

[1] Division of Biomedical & Life Sciences, Faculty of Health & Medicine, Lancaster University, Lancaster LA1 4YT, UK.

[2] Department of Mathematics & Statistics, Lancaster University, Lancaster LA1 4YF, UK.

* Corresponding author: d.gatherer@lancaster.ac.uk, +44 1524 592900, Twitter: @DerekGatherer

Abstract

The protein sequence-structure gap results from the contrast between rapid, low-cost deep sequencing, and slow, expensive experimental structure determination techniques. Comparative homology modelling may have the potential to close this gap by predicting protein structure in target sequences using existing experimentally solved structures as templates. This paper presents the first use of force-directed graphs for the visualization of sequence space in two dimensions, and applies them to the choice of suitable RNA-dependent RNA polymerase (RdRP) target-template pairs within human-infective RNA virus genera. Measures of centrality in protein sequence space for each genus were also derived and used to identify centroid nearest-neighbour sequences (CNNs) potentially useful for production of homology models most representative of their genera. Homology modelling was then carried out for target-template pairs in different species, different genera and different families, and model quality assessed using several metrics. Reconstructed ancestral RdRP sequences for individual genera were also used as templates for the production of ancestral RdRP homology models. High quality ancestral RdRP models were consistently produced, as were good quality models for target-template pairs in the same genus. Homology modelling between genera in the same family produced mixed results and inter-family modelling was unreliable. We present a protocol for the production of optimal RdRP homology models for use in further experiments, e.g. docking to discover novel anti-viral compounds. (219 words)

Keywords: *Homology modelling, protein structure, RNA-dependent RNA polymerase, reverse transcriptase, sequence- structure gap, sequence space, Fruchterman-Reingold algorithm, force-directed graphs, multidimensional scaling.*

1. Introduction

Since high-throughput sequencing technologies entered mainstream use towards the end of the first decade of the 21st century, there has been an explosion in available protein sequences. By contrast, there has been no corresponding high-throughput revolution in structural biology. Obtaining solved structures of proteins at adequate resolution remains a painstaking task. X-ray crystallography is still the gold standard for structure determination more than 60 years after its first use in determining myoglobin structure [1]. The result of this discrepancy between the rate of protein sequence determination and the rate of protein structure determination is the *protein sequence-structure gap* [2].

Homology modelling is a rapid computational technique for prediction of a protein's structure from a) the protein's sequence, and b) a solved structure of a related protein, referred to as the target and the template, respectively. Since structural similarity often exists even where sequence similarity is low [2, 3], homology modelling has the potential to reduce massively the size of the protein sequence-structure gap, provided the models produced can be considered reliable enough for use in further research.

The RNA-dependent RNA polymerase (RdRP) of RNA viruses presents an opportunity to test and expand this approach. RdRPs are the best conserved proteins throughout the RNA viruses, being essential for their replication [4]. Conservation is particularly high in structural regions that are involved in the replication process, for instance the indispensable RNA-binding pocket [5]. RdRPs are also of immense medical importance as the principal targets for anti-viral drugs. Evolution of resistance against anti-viral drugs is a major concern for the future, and the design of novel anti-viral compounds is a highly active research area. Solved structures of RdRPs are of great assistance to these efforts, as they enable the use of docking protocols against large libraries of pharmaceutical candidate compounds [e.g. 6, 7].

Although some human-infective RNA viruses have solved RdRP structures, there are still large areas within the virus taxonomy that lack any. This paper will first identify where the protein sequence-structure gap is at its widest in RdRPs. Because of the sequence-structure gap, it is therefore impossible in many genera to perform docking protocols against solved structures of RdRP for discovery of novel anti-viral compounds. Under these circumstances, replacement of real solved structures with homology models for docking experiments requires that the homology models used should be both high quality and also optimally representative of their respective genera. Our

second task is to present several similarity metrics in sequence space that assist in the identification of the virus species having the RdRP sequence that is most representative of its genus as a whole. We then present the first use of force-directed graphs to produce an intuitive visualization of sequence space, and select target RdRPs without solved structures for homology modelling. These are then used to perform homology modelling using template-target pairs within the same genus, between sister genera and between sister families, monitoring the quality of the models produced as the template becomes progressively more genetically distant to the target sequence being modelled. Finally, we produce homology models for reconstructed common ancestral RdRP sequences. In the light of our results, we comment on the strengths and weakness of homology modelling to reduce the size of the protein sequence-structure gap for RdRPs, and produce a flowchart of recommendations for docking experiments on RdRP proteins lacking a solved structure.

2. Materials & Methods

2.1 Taxonomy search

We chose RdRPs from human-infective viruses based on the list provided by Woolhouse & Brierley [8]. Given the global medical importance of AIDS, we also included *Lentivirus* reverse transcriptases (RTs) for analysis. Solved structures for these proteins, where available, were downloaded from the RCSB Protein Data Bank (PDB) [9]. Table 1 presents our criteria for selecting suitable homology modelling candidates.

2.2 Multiple sequence alignment

RdRP and RT amino acid sequences for all virus species satisfying the criteria of Table 1 were downloaded from GenBank [10]. Alignment of sequence sets for each genus, was performed using MAFFT [11]. Alignments were refined in MEGA [12] using Muscle [13] where necessary, and the best substitution model determined. Alignment of target sequences onto their solved structure templates for homology modelling was carried out using the Molecular Operating Environment (MOE v.2016.08, Chemical Computing Group, Montreal H3A 2R7, Canada).

2.3 Visualization of sequence space

We define sequence space as a theoretical multi-dimensional space within which protein sequences may be represented by points. For an alignment of *N* related proteins, the necessary

dimensionality of this sequence space is *N*-1, with the hyperspatial co-ordinates in each dimension for any protein determined by its genetic distance to the *N*-1 other proteins. For *N* = 5, direct visualization of all dimensions of sequence space is impractical at best, since a 4-dimensional space must be simulated in three dimensions, and is effectively impossible for N ≥ 6. The following methods were used to reduce sequence space to two and three dimensions for ease of visualization. To simplify calculations, we allow an extra dimension defined by the distance from each sequence to itself. The value of the co-ordinate in that dimension is always zero and our sequence space has *N* dimensions rather than *N*-1.

2.3.1 Two-dimensional visualization of sequence space

The pairwise distance matrix (M_d) for each genus, calculated from the sequence alignment in MEGA, consists of entries $M_d(i,j)$ giving the genetic distance between each pair of sequences *i* and *j* where $\{i, j\} \in \{1,2 \ldots .. N\}$ and $i \neq j$, for a set of *N* sequences. In our data set *N* ranges (see Supplementary Table) from 5 (genus *Picobirnavirus*) to 64 (genus *Flavivirus*).

For each alignment, the pairwise distance matrix (M_d) was converted into a similarity matrix (M_s) as follows:

$$M_s(i,j) = 1/(M_d(i,j) + 1) \quad (1)$$

The similarity matrix was then used as input for R package *qgraph* [14]. The "spring" layout option was chosen, which uses the Fruchterman-Reingold algorithm to produce a two-dimensional undirected graph in which edge thickness is proportional to absolute distance in *N* dimensions and node proximity in two dimensions is optimized for ease of viewing while attempting to ensure that those nodes closely related in the *N*-dimensional input are also close in the two-dimensional output [15]. 500 iterations were performed, or until convergence was achieved.

2.3.2 Three-dimensional visualization of sequence space

For each alignment, the pairwise distance matrix (M_d) was used as input for R package *cmdscale*, which uses multi-dimensional scaling to produce a three-dimensional graph from the *N*-dimensional input, with node proximity again reflecting relative similarity [16]. Spotfire Analyst (TIBCO Spotfire Analyst, v.7.12.0, 2018) was used to visualize the output of *cmdscale*.

2.4 Centroid nearest neighbour determination

We define the *centroid* as a hypothetical protein sequence located at the centre point of the sequence space of an alignment. The real sequence closest to the hypothetical centroid is termed the *centroid nearest neighbour* (CNN). We calculate the position of the CNN in three ways.

2.4.1 Shortest-path centroid nearest neighbour

For a sequence $i \in \{1,2 \ldots\ldots N\}$ in an alignment of N sequences, its total path length $D(i)$ to the other N-1 sequences may be calculated from the distance matrix M_d as follows:

$$D(i) = \sum_{j=1}^{j=N} M_d(i,j) \tag{2}$$

where $i = j$, $M_d(i,j)$ is zero. This may be omitted to enforce a strict N-1 dimensions for N input sequences, but we leave it in to simplify subsequent calculations. We define i^* as the index that minimizes $D(i)$.

$$D(i^*) = \underset{1 \le i \le N}{\operatorname{argmin}} \sum_{j=1}^{j=N} M_d(i,j) \tag{3}$$

The shortest path CNN is therefore sequence i^*. For alignments where clusters of closely related sequences exist, giving many values of $M_d(i,j)$ close to zero, this method will tend to place the CNN within a cluster. To overcome this problem, the arithmetic mean and median, respectively, were used to determine the mean CNN and the median CNN.

2.4.2 Mean centroid nearest neighbour

The values of D (equation 2) may be averaged to produce mean total path distance $\bar{D}$:

$$\bar{D} = \left(\sum_{i=1}^{i=N} D(i) \right) \Big/ N \tag{4}$$

where again N is the total number of sequences in the alignment. We now re-define i^* as the index that minimizes $D(i) - \bar{D}$.

$$D(i^*) = \underset{1 \le i \le N}{\operatorname{argmin}}(D(i) - \bar{D}) \tag{5}$$

In the event of equation 5 returning zero, the mean CNN and the true centroid are identical. As with all variables using means, the mean CNN is liable to skewing by outliers.

2.4.3 Median centroid nearest neighbour

We generate a vector D over $i \in \{1,2 \ldots\ldots N\}$, in which each entry $D(i)$ represents the total path length for sequence i (equation 2). The values of vector D are then ranked in ascending order $x_{\sigma(1)}$ to $x_{\sigma(N)}$ to produce vector D_σ.

$$D_\sigma = \left\{D\left(i, x_{\sigma(1)}\right), D\left(i, x_{\sigma(2)}\right) \ldots D\left(i, x_{\sigma(N)}\right)\right\} \tag{6}$$

The median CNN is the sequence with value $D(i)$ situated in the middle of the array D_σ, at $D(m)$, where $D(m)$ is either $D(m_{odd})$ or $D(m_{even})$ for alignments with odd or even numbers of sequences respectively.

$$D(m_{odd}) = D\left(i, x_{\sigma((N+1)/2)}\right) \tag{7}$$

$$D(m_{even}) = \left(D\left(i, x_{\sigma(N/2)}\right) + D\left(i, x_{\sigma((N/2)+1)}\right)\right)/2 \tag{8}$$

We now re-define i^* as the index that minimizes $D(i) - D(m)$.

$$D(i^*) = \underset{1 \le i \le N}{\operatorname{argmin}}(D(i) - D(m)) \tag{9}$$

Again, in the event of equation 9 returning zero, the median CNN and the true centroid are identical. As with all variables using medians, the median CNN is liable to skewing by the presence in the alignment of multiple sequences with the same value of *D(i)*.

2.5 Homology modelling

The choice of solved structures as templates for homology modelling, and the choice of targets to be modelled, within each genus was governed by the following rules:

1) For each genus the solved structure that covered the highest proportion of the RdRP or RT sequence was chosen as the template for that genus.
2) If more than one candidate template structure was found at this sequence length, the structure with the lowest resolution in angstroms was selected. See Table 2 for the templates satisfying these two criteria.
3) Within each genus, the sequence with the greatest genetic distance from the template, was chosen as the target for homology modelling. See Table 3 for the template-target pairs satisfying this criterion.
4) Criterion 3 was applied to find template-target pairs in different genera (see Table 4) and different families (see Table 5), thus testing the limits of homology modelling at high genetic distances.

Homology modelling was carried out using the Molecular Operating Environment (MOE v.2016.08, Chemical Computing Group, Montreal H3A 2R7, Canada). Ten intermediate models were produced using the Amber10:EHT forcefield under medium refinement. The model that scored best under the generalised Born/volume integral (GB/VI) was selected to undergo further energy minimisation using Protonate3D, which predicts the location of hydrogen atoms using the model's 3D coordinates [17, 18].

2.6 Model quality analysis

2.6.1 Φ-Ψ outliers

To assess the stereochemical quality of the homology models produced, Ramachandran plots were derived in MOE, and used to calculate the proportion of bad outlier Φ-Ψ angles in the model, after subtraction of the number of outlier Φ-Ψ angles in the template. Generally, outlier angle percentage below 0.05% indicates a very high quality model, and a percentage below 2%

indicates a good quality model [19].

2.6.2 Root-mean-square deviation

Models were superposed with their templates in MOE and root-mean-square deviation (RMSD) value derived for the alpha carbons (Cα) in the two structures. Generally, an RMSD below 2 Å indicates a good quality model [20].

2.6.3 QMEAN Z-score

Qualitative Model Energy Analysis (QMEAN) was used to analyse models using both statistical and predictive methods [21]. The QMEAN Z-score is an overall measure of the quality of the model when compared to similar models from a PDB reference set of X-ray crystallography-solved structures. A Z-score of 0 would indicate a model of the same quality as a similar high quality X-ray crystallographic structure, while a Z-score below -4.00 indicates a low quality model [22].

2.7 Ancestral sequence reconstruction and modelling

Maximum likelihood (ML) trees [23] were produced for each genus in MEGA. The ML tree and the corresponding multiple sequence alignment were input into the ancestral reconstruction server, FASTML [24]. The reconstructed sequence for the root of the tree, i.e. the putative common ancestor RdRP or RT sequence for the genus was used as the target for homology modelling in MOE, using the template chosen according to the rules in section 2.5. The reconstructed ancestral sequence was added to the alignment and the force-directed graph re-drawn. Figure 1B, showing the target-template pairs for homology modelling may be compared with Figure 1C, showing the ancestor-template pairs.

3. Results

3.1 Areas of the taxonomy that lack solved RdRP structures

Our first observation is that there are still large areas of the viral taxonomy where no solved RdRP structures exist. No suitable templates for homology modelling were found within the entire *Nidovirales* order of RNA viruses. This order contains several coronaviruses important to human health including *Severe acute respiratory syndrome-related coronavirus* (SARS-CoV) and *Middle East respiratory syndrome-related coronavirus* (MERS-CoV) [25]. In the order *Mononegavirales*, *Vesiculovirus* was the only genus with a solved RdRP structure suitable for

homology modelling. However, this order contains many medically important viruses such as *Zaire ebolavirus*, *Hendra henipavirus*, *Measles morbillivirus*, and *Mumps rubulavirus* [26]. In the order *Bunyavirales*, *Phenuiviridae* stands out as an important family lacking a solved RdRP, despite it containing various human-infective arboviruses such as *Rift Valley fever phlebovirus* and *Sandfly fever Naples phlebovirus* [27].

Furthermore, some genera have solved RdRP structures which only cover a small proportion of the protein. For instance, *Orthohantavirus*, *Orthonairovirus* and *Mammarenavirus* only have solved structures covering less than 10% of the RdRP sequence (Table 1).

3.2 Sequence space visualization

3.2.1 Two-dimensional visualisation

Figure 1 shows two-dimensional force-directed graphs of similarity for each genus with more than four RdRP reference sequences (or RT sequences in the case of *Lentivirus*). In principle, it would be possible to draw force-directed graphs for entire families and even orders. However, the input to *qgraph* is the similarity matrix calculated from the distance matrix, and the distance matrix is calculated in MEGA from an alignment. Once taxonomic distance begin to extend beyond genera, alignment becomes progressively less reliable, with all the downstream statistics tending to degrade as a consequence. We therefore confine our construction of force-directed graphs to intra-genus comparisons.

It is evident from Figure 1 that sequences are not necessarily evenly distributed in sequence space. Clustering is noticeable in the genus *Flavivirus*, with two sub-groups and an outlier sequence evident. *Mammarenavirus* also shows division into two sub-groups. By contrast, *Picobirnavirus* has only five relatively equidistant reference sequences, thus producing a highly regular pentagram. Similarly, *Rotavirus* has eight reference sequences, with four at each end of a fairly regular cuboid. Figure 1A also shows how the various methods (equations 2-9) for determining the CNN of sequence space for each genus, are in poor agreement. Only in *Rotavirus* and *Picobirnavirus* are mean and median CNNs found in the same sequence. Figure 1A also shows that the best solved structure for the purposes of template choice in homology modelling is rarely close to the centre of sequence space. Only in *Lentivirus* is the optimal template also the mean CNN, and only in *Vesiculovirus* is the optimal template a shortest-path CNN. Figure 1B shows the relations of the template-target pairs in sequence space, illustrating how intra-genus homology modelling template-target selection attempts to traverse the largest genetic distance available within the genus.

3.2.2 Three-dimensional visualisation

Figures 2 and 3 compare, for genera *Orthohantavirus*,and *Mammarenavirus* respectively, the force-directed graphs of Figure 1 with the three-dimensional equivalent output of multidimensional scaling. Figure 2 shows a sequence clustering within *Orthohantavirus* that is not readily apparent in the force-directed graph. The CNNs are distributed among four clusters, as there is no sequence close to the geometrical centre of the three-dimensional space, where the notional centroid is located. The solved structure has 10 other sequences in its proximity in the three-dimensional space, roughly equivalent to the lower right quadrant of the two-dimensional force-directed graph. Similarly, the shortest-path CNN and mean CNN are both located are located within another three-dimensional cluster also containing 11 sequences, which is roughly equivalent to the upper right quadrant of the two-dimensional force-directed graph.

Figure 3 presents a similar picture for *Mammarenavirus*. The force-directed graph for *Mammarenavirus* has more obvious clustering that that for *Orthohantavirus*, showing a lower-left to top-right split. In the three-dimensional representation, these are equivalent, respectively, to the three clusters on the right and two clusters on the left. As with *Orthohantavirus*, there is no CNN near the geometrical centre of the three-dimensional space, but the CNNs are distributed around two clusters.

Three dimensional representations of all the genera in Figure 1 are available from the link in the Raw Data section.

3.3 Homology modelling

Homology modelling was carried out as follows:

1) Intra-genus, inter-species (11 models, Table 3)
2) Intra-family, inter-genus (5 models, Table 4)
3) Intra-order, inter-family (7 models, Table 5)
4) Intra-genus, on reconstructed common ancestor (12 models, Table 6)

Table 3 shows that homology modelling with template and target within the same genus, produced good quality models in most cases, as judged by percentage of Φ-Ψ outliers and RMSD within the high quality range. Only the models for *American bat vesiculovirus* and *Tamana bat virus* have percentages of Φ-Ψ outliers outside of the high quality range. QMEAN, however, is rather more critical of the output with only the model for *Porcine picobirnavirus* falling within

the high quality range. The model for *Imjin thottimvirus* scores eighth best on percentage of Φ-Ψ outliers and second best on RMSD, despite the re-classification (occurring after the completion of our experimental work) by the ICTV of this virus, originally in genus *Orthohantavirus* into a new *Thottimvirus* genus [28]. It should be noted that the models for *Imjin thottimvirus*, *Burana orthonairovirus* and *Brazilian mammarenavirus* were based on very short template structures (see Table 2).

Table 4 shows that homology modelling with template and target within the same family but different genera, still produced good quality models in most cases, as judged by percentage of Φ-Ψ outliers and RMSD within the high quality range. Only the models for *Lleida bat lyssavirus* and *Macaque simian foamy virus* have percentages of Φ-Ψ outliers outside of the high quality range. However, once again, QMEAN assesses all models as outside the high quality range.

Table 5 shows that homology modelling with template and target within the same order but in different families, is a far more difficult proposition than at the lower taxonomic levels. The model for *Mammalian orthobornavirus 1* fails all three quality tests and only the model for *Rift Valley fever phlebovirus* manages to pass two out of three.

Table 6 shows that modelling the structure of the reconstructed sequence of the common ancestor of each genus, produces models of the same standard as intra-genus modelling (compare Tables 3 and 6). By contrast with almost all the other models, the QMEAN scores are within the high quality range, with only two exceptions, the common ancestors of genera *Rotavirus* and *Vesiculovirus*. Figure 1C shows the force-directed graphs with the locations of the ancestral sequences added.

Table 7 summarises the results of Tables 3 to 6 inclusive. As the taxonomical distance increases, production of high quality homology models becomes more difficult. However, modelling the reconstructed ancestral sequence of each genus is typically productive of a better scoring model even than the real sequence targets chosen for intra-genus modelling.

Figure 4 shows representative examples of homology models of high and low quality superimposed with their template solved structure along with their corresponding Ramachandran plot and QMEAN quality scores.

All homology models in Tables 3 to 6 are available from the link in the Raw Data section.

4. Discussion

The first objective of this study was to identify viral taxa which are comparatively lacking in solved structures for RNA-dependent RNA polymerase (RdRP). We observed that the entire order *Nidovirales*, the families *Bornaviridae*, *Filoviridae* and *Paramyxoviridae* within the order *Mononegavirales*, and the family *Phenuiviridae* within the order *Bunyavirales*, fall into this category. Additionally, within the genera *Orthohantavirus*, *Orthonairovirus* and *Mammarenavirus*, all within the order *Bunyavirales*, the solved structure available for RdRP covers less than 10% of the protein sequence. Given the medical importance of many viruses within these taxa, and the number of anti-viral drugs that target RdRPs we suggest that they are prioritized for X-ray crystallography to close the "sequence-structure gap".

Our second objective was to assess how well homology modelling could provide models that might serve for computer-assisted drug discovery of novel anti-viral compounds. To assist in the visualization of sequence space, we produced the first application of force-directed graphs to protein sequences (Figure 1). We also applied multidimensional scaling for comparative purposes (Figures 2 and 3). Force-directed graphs enable the visualization of complex data in two dimensions. The three dimensional visualization produced from multidimensional scaling is visually richer, but this benefit can only be appreciated when a viewing application such as Spotfire is available so that the three-dimensional image can be rotated. Force-directed graphs convey much of the information in a single image which may be printed on a page or viewed on screen. This two-dimensional collapsing of sequence space also allows for easy simultaneous comparison of multiple datasets, in the present case multiple genera, which cannot readily be performed if separate three-dimensional viewers require to be open.

The most common method of visualizing sequence space is the phylogenetic tree. For instance, starting from a distance matrix, agglomerative hierarchical clustering, such as the UPGMA method [29], can be performed to generate a tree. Slightly more sophisticated methods, such as neighbour-joining [30] can generate trees where the branch lengths are proportional to genetic distance. Force-directed graphs do not represent genetic distance as accurately as phylogenetic trees, since the distances between nodes, although optimized to reflect relatedness, are constrained by the Fruchterman-Reingold algorithm to the best representation in two dimensions. However, force-directed graphs again allow easier simultaneous comparison of several data sets than phylogenetic trees. Figure 1 would be impossible to create on a single page if trees were used instead of force-directed graphs. Trees represent ancestral sequences as nodes on the tree, with only existing taxa as leaves. Force-directed graphs, by contrast, allow ancestral sequences to be represented in the same way as existing ones. Figure 1C shows that ancestral sequences do

not necessarily appear as outliers in force-directed graphs. Indeed, for genera *Flavivirus*, *Hepacivirus*, *Orthobunyavirus* and *Orthohantavirus* in particular, the insertion of the reconstructed ancestral sequence into the force-directed graph in Figure 1C does not overly distort its original shape in Figures 1A and 1B. The reason for this becomes apparent when one considers a phylogenetic tree represented in unrooted "star" format. The ancestral sequence is then at the centre of the star topology and it can be seen that the genetic distance from the root to any particular leaf sequence may often be less than for many pairwise leaf sequence combinations. We did not perform calculation of centroid nearest neighbours (CNNs) for alignments incorporating reconstructed ancestral sequences, but we are tempted to speculate that many of the ancestral sequences would have been CNNs, had they been included.

It is important to remember that homology models are theoretical constructions and caution must be exercised in treating them as input material for further experiments. Among the various statistics for assessment of model quality, Φ-Ψ outlier percentage is a measure of the proportion of implausible dihedral angles in the model, and indicate where parts of the model backbone are likely to be incorrectly predicted. Nevertheless, it is also important not to become too dependent on statistics such as Φ-Ψ outlier percentage, as "bad" angles do occasionally occur in solved structures. For instance in the present study, the thresholds of < 0.05% for a very high quality model, and < 2% for a good quality model given by Lovell *et al* [19] would suggest that six of the twelve template solved structures used here (Table 2) would not have been assessed as "very high quality" had they been models rather than solved structures. Indeed the templates from *Indiana Vesiculovirus* and *Rotavirus A* have more than 0.5% Φ-Ψ outliers, and also have the poor quality scores for QMEAN. These two structures also have the poorest resolution of any of our templates, at > 3Å. The poor quality scoring may therefore simply be a consequence of uncertainties in positioning of atoms in these structures. One might reasonably posit that the use of template solved structures having such issues might influence the resulting models to contain the same outliers. However, the model for *Rotavirus I* has a lower level of Φ-Ψ outliers than its *Rotavirus A* template (Table 3).

As might be expected, production of high quality models becomes more difficult as the genetic distance between target and template increases, as show in Tables 3 to 5. Nevertheless, even at the level of template-target pairs in separate genera (Table 4), the average performance is acceptable, as summarized in Table 7. We therefore suggest that homology modelling may be used to produce RdRP models for research use even for genera where no solved structure exists, provided a template structure exists within the same family. Here, we provide examples (Table 4) of such successful inter-genus, intra-family, models for genera *Coltivirus* and *Parechovirus*. Our inter-genus models for *Lyssavirus* and *Spumavirus* are slightly less successful. Moving to the next taxonomic level, models

with template-target pairs in separate families (Table 5) are generally less successful. One exception is our model for family *Phenuiviridae*, which is better than some of the intra-family models. This is encouraging, since *Phenuiviridae* is a family without any solved RdRP structure. Homology models have been produced at much larger taxonomic distances than those dealt with here, for instance from bacteria to eukaryotes [31], so it should be stressed that we make no claim for the generality of our findings outside of the viral orders under consideration, or for proteins other than RdRP. Multi-domain proteins in particular, may produce higher quality models for some domains than others.

One surprising result was the high quality of the models of reconstructed ancestral sequences (Table 6, summarized in Table 7). As previously discussed, this may be due to the fact that the ancestral sequence is, assuming a regular molecular clock, potentially equally related to all descendent members of its genus. In this paper, we calculated centroid nearest neighbours (CNNs) as the central points in sequence space for each genus (Figure 1). A reconstructed ancestral sequence may also be considered as a candidate central point. The value of central points is that they may serve as targets that could be used to make models representative of their genus as a whole. For instance, the shortest-path, mean and median CNNs of genus *Orthohantavirus* are sequences 16, 22 and 7 (see Supplementary Table for a list of sequences for each genus), representing *Sin Nombre orthohantavirus*, *Rockport orthohantavirus* and *Cao Bang orthohantavirus* respectively. The partial solved structure used as the template for modelling in the genus *Orthohantavirus* in the present paper is from *Hantaan orthohantavirus* (5IZE, see Table 2) and the target used, *Imjin thottimvirus* (sequence 27 in *Orthohantavirus* panel of Figure 1), is now classified as belonging to a new genus *Thottimvirus* (Table 3). The three CNNs, *Sin Nombre orthohantavirus*, *Rockport orthohantavirus* and *Cao Bang orthohantavirus* are 71%, 64% and 75% identical to 5IZE respectively, whereas *Imjin thottimvirus* is only 58% identical. The latter was of course chosen to test the effectiveness of intra-genus homology modelling over as wide a genetic distance as possible (see Section 2.5). For the performance of subsequent experimental procedures on *Orthohantavirus* RdRPs, for instance docking to discover novel anti-viral compounds, a homology model corresponding to one of the three CNNs mentioned above or to the reconstructed ancestor (Table 6) would be the preferred target, along with the existing solved structure.

On the basis of our investigations, we recommend a procedural flowchart for selection of an RdRP structure for further study, for instance docking to discover novel anti-viral compounds, in any RNA virus genus of interest (Figure 5). Where a solved structure exists within a genus, it is the obvious choice for further experiments. However, where that solved structure is far from any of the CNN sequences of the genus, as judged by the force-directed graph, a CNN may also be homology modelled for comparative purposes, using the existing solved structure as a template. Any differential performance of the solved structure and the homology model in, for instance, a docking experiment,

may give clues as to the generality of conclusions derived from the solved structure alone. A reconstructed ancestral RdRP may also be used as an alternative to, or in addition to, a CNN. The limits of homology modelling would appear, on the basis of the results presented here, to be at the intra-family, inter-genus level. Template-target pairs in different viral families are unlikely to be of practical use, as the predicted quality of the resulting models is low. Our models were produced using MOE, and we have not performed comparisons using other modelling tools, such as SWISS-MODEL[31] or Modeller [32]. We feel that it is unlikely that significant differences in output would be produced, but when the object of the exercise is drug-discovery, we recommend that the protocol in Figure 5 be implemented using several alternative modelling softwares.

Crystallographic structural genome projects are badly needed to close the sequence-structure gap. In the meantime, systematic attempts to fill the gaps via homology modelling may be useful. However, for many taxa – all of the order *Nidovirales* and much of *Mononegavirales* - the paucity of solved structures to act as templates remains a serious obstacle.

Raw Data

All code, inputs and outputs are available from:
https://doi.org/10.17635/lancaster/researchdata/276

References

[1] Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., Phillips, D.C. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature. 1958, 181, 662-6.
[2] Schwede, T. Protein modeling: what happened to the "protein structure gap"? Structure. 2013, 21, 1531-40.
[3] Kaczanowski, S., Siedlecki, P., Zielenkiewicz, P. The High Throughput Sequence Annotation Service (HT-SAS) - the shortcut from sequence to true Medline words. BMC Bioinformatics. 2009, 10, 148.
[4] Holmes, E.C. The Evolution and Emergence of RNA Viruses. Oxford, UK, Oxford University Press, 2009.
[5] Lu, G., Gong, P. Crystal Structure of the full-length Japanese encephalitis virus NS5 reveals a conserved methyltransferase-polymerase interface. PLoS Pathog. 2013, 9, e1003549.
[6] Elfiky, A.A., Ismail, A.M. Molecular docking revealed the binding of nucleotide/side inhibitors to Zika viral polymerase solved structures. SAR QSAR Environ Res. 2018, 29, 409-18.
[7] Nncube, N.B., Ramharack, P., Soliman, M.E.S. Using bioinformatics tools for the discovery of Dengue RNA-dependent RNA polymerase inhibitors. PeerJ. 2018, 6, e5068.
[8] Woolhouse, M.E.J., Brierley, L. Epidemiological characteristics of human-infective RNA viruses. Sci Data. 2018, 5, 180017.

[9] Rose, P.W., Prlic, A., Altunkaya, A., Bi, C., Bradley, A.R., Christie, C.H., et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res. 2017, 45, D271-D81.
[10] O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016, 44, D733-45.
[11] Katoh, K., Standley, D.M. MAFFT: iterative refinement and additional methods. Methods Mol Biol. 2014, 1079, 131-46.
[12] Kumar, S., Stecher, G., Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016, 33, 1870-4.
[13] Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004, 32, 1792-7.
[14] Epskamp, S., Cramer, A., Waldorp, L., Schmittmann, V., Borsboom, D. qgraph: Network visualizations of of relationships in psychometric data. Journal of Statistical Software. 2012, 48, 1-18.
[15] Fruchterman, T.M.J., Reingold, E.M. Graph drawing by force-directed placement. Software – Practice & Experience. 1991, 21, 1129–64, doi:10.002/spe.4380211102.
[16] Mardia, K.V. Some properties of classical multidimensional scaling. Communications on Statistics – Theory and Methods. 1978, A7, 1233–41. doi: 10.080/03610927808827707.
[17] Labute, P. Protonate3D: assignment of ionization states and hydrogen coordinates to macromolecular structures. Proteins. 2009, 75, 187-205.
[18] Labute, P. The generalized Born/volume integral implicit solvent model: estimation of the free energy of hydration using London dispersion instead of atomic surface area. J Comput Chem. 2008, 29, 1693-8.
[19] Lovell, S.C., Davis, I.W., Arendall, W.B., 3rd, de Bakker, P.I., Word, J.M., Prisant, M.G., et al. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins. 2003, 50, 437-50.
[20] Forrest, L.R., Tang, C.L., Honig, B. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. Biophys J. 2006, 91, 508-17.
[21] Benkert, P., Tosatto, S.C., Schomburg, D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins. 2008, 71, 261-77.
[22] Benkert, P., Biasini, M., Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics. 2011, 27, 343-50.
[23] Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981, 17, 368-76.
[24] Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., et al. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res. 2012, 40, W580-4.
[25] de Wit, E., van Doremalen, N., Falzarano, D., Munster, V.J. SARS and MERS: recent insights into emerging coronaviruses. Nat Rev Microbiol. 2016, 14, 523-34.
[26] Maes, P., Amarasinghe, G.K., Ayllon, M.A., Basler, C.F., Bavari, S., Blasdell, K.R., et al. Taxonomy of the order Mononegavirales: second update 2018. Arch Virol. 2019, 164, 1233-44.
[27] Elliott, R.M., Brennan, B. Emerging phleboviruses. Curr Opin Virol. 2014, 5, 50-7.
[28] Maes, P., Adkins, S., Alkhovsky, S.V., Avsic-Zupanc, T., Ballinger, M.J., Bente, D.A., et al. Taxonomy of the order Bunyavirales: second update 2018. Arch Virol. 2019, 164, 927-41.
[29] Prager, E.M., Wilson, A.C. Construction of phylogenetic trees for proteins and nucleic acids: empirical evaluation of alternative matrix methods. J Mol Evol. 1978, 11, 129-42.
[30] Saitou, N., Nei, M. The neighbor-joining method: a new method for reconstructing

phylogenetic trees. Mol Biol Evol. 1987, 4, 406-25.
[31] Guex, N., Peitsch, M.C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis. 1997, 18, 2714-23.
[32] Fiser, A., Sali, A. Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol. 2003, 374, 461-91.

Criterion	**Reason**
Human-infective virus	Importance to human health
NCBI RefSeq annotated genome	Easy retrieval of high quality RdRP sequence
RdRP located at the 3' end of polyprotein or on its own segment	Eliminates unconventional RdRPs
At least one solved RdRP at a range of different taxonomic levels, e.g. in same species, same genus, same family, same order.	To be used as the templates in homology modelling at different levels of genetic distance.

Table 1: List of criteria used to select RNA-dependent RNA polymerases (RdRPs) for homology modelling.

Order	Family	Genus	Species	PDB	Resolution (Å)	RdRP coverage (%)	Φ-Ψ outliers (%)	QMEAN Z-score	Reference
Bunyavirales	*Hantaviridae*	*Orthohantavirus*	*Hantaan orthohantavirus*	5IZE	1.70	8	0.00	0.36	Reguera et al 2016
	Nairoviridae	*Orthonairovirus*	*Crimean-Congo hemorrhagic fever orthonairovirus*	3PHX	1.60	5	0.00	0.72	Akutsu et al 2011
	Peribunyaviridae	*Orthobunyavirus*	*La Crosse orthobunyavirus*	5AMQ	3.00	100	0.06	-1.60	Gerlach et al 2015
	Arenaviridae	*Mammarenavirus*	*Lymphocytic choriomeningitis mammarenavirus*	3JSB	2.13	9	0.00	-0.31	Love et al 2014
Mononegavirales	*Rhabdoviridae*	*Vesiculovirus*	*Indiana vesiculovirus*	5A22	3.80	100	0.95	-5.11	Morin et al 2010
Picornavirales	*Picornaviridae*	*Enterovirus*	*Rhinovirus A*	1XR7	2.30	>99	0.00	-0.16	Vives-Adrian et al 2014
		Cardiovirus	*Cardiovirus A*	4NYZ	2.15	100	0.22	0.00	Liang et al 2015
No order assigned	*Flaviviridae*	*Flavivirus*	*Japanese encephalitis virus*	4K6M	2.60	100	0.00	-0.91	Lu & Gong 2013
		Hepacivirus	*Hepacivirus C*	2YOJ	1.76	>98	0.00	0.32	Chen et al 2014
	Picobirnaviridae	*Picobirnavirus*	*Human picobirnavirus*	5I61	2.40	100	0.19	-0.75	Collier et al 2016
	Reoviridae	*Rotavirus*	*Rotavirus A*	2R7O	3.35	100	1.37	-4.35	Lu et al 2008
Ortervirales	*Retroviridae*	*Lentivirus*	*Human immunodeficiency virus 1*	5TXL	2.50	100	0.18	-0.61	Das et al 2017

Table 2: **Solved structures of RdRPs and reverse transcriptase (for HIV-1) selected as templates for homology modelling.** All are derived by X-ray crystallography except 5A22 which is a cryo-electron microscopy structure. For protein coverage, blue indicates that the template covers more than 90% of the

sequence, red indicates less. For Φ-Ψ outliers and QMEAN Z-score, blue indicates good-quality, red indicates poor-quality, determined by the following thresholds: Φ-Ψ = 2%, QMEAN Z-score = -4.00

Genus	Template species	Template PDB	Target species	Target reference genome	Φ-Ψ outliers (%)	RMSD (Å)	QMEAN Z-score
Orthohantavirus	*Hantaan orthohantavirus*	5IZE	*Imjin thottimvirus* *	NC_034564	1.67	0.499	-4.12
Orthonairovirus	*Crimean-Congo hemorrhagic fever orthonairovirus*	3PHX	*Burana orthonairovirus* (*Tacheng tick virus*)	NC_031284	0.00	1.222	-4.74
Orthobunyavirus	*La Crosse orthobunyavirus*	5AMQ	*Shuni orthobunyavirus* (*Aino virus*)	NC_018465	0.87	1.175	-4.02
Vesiculovirus	*Indiana vesiculovirus*	5A22	*American bat vesiculovirus*	NC_022755	3.22	1.007	-10.27
Enterovirus	*Rhinovirus A*	1XR7	*Enterovirus E*	NC_001859	1.52	0.564	-4.83
Mammarenavirus	*Lymphocytic choriomeningitis mammarenavirus*	3JSB	*Brazilian mammarenavirus* (*Sabia virus*)	NC_006313	1.73	0.401	-4.97
Flavivirus	*Japanese encephalitis virus*	4K6M	*Tamana bat virus*	NC_003996	2.06	1.191	-5.80
Hepacivirus	*Hepacivirus C*	2YOJ	*Hepacivirus N*	NC_038432	1.20	0.861	-4.42
Picobirnavirus	*Human picobirnavirus*	5I61	*Porcine picobirnavirus*	NC_029802	1.33	0.586	-3.98
Rotavirus	*Rotavirus A*	2R7O	*Rotavirus I*	NC_026825	0.42	0.949	-6.54
Lentivirus	*Human immunodeficiency virus 1*	5TXL	*Caprine arthritis encephalitis virus*	NC_001463	0.55	0.778	-4.01

Table 3: Homology modelling at intra-genus, inter-species level. Templates are as given in Table 2. Targets are the RdRP (or reverse transcriptase for *Lentivirus*) sequences from the reference genome accession numbers given. RMSD: root mean square deviation in Angstroms between template and model when superposed in MOE. Blue indicates good quality, red indicates poor quality, determined by the following

thresholds: Φ-Ψ < 2%; QMEAN Z-score > -4.00; RMSD < 2 Å. Purple indicates good quality, but using a partial template (see Table 1) *_Imjin thottimvirus_ was reclassified in 2018 by the International Committee on Taxonomy of Viruses (ICTV) in a new genus _Thottimvirus_

Family	Template species	Template PDB	Template genus	Target genus	Target species	Target reference genome	Φ-Ψ outliers (%)	RMSD (Å)	QMEAN Z-score
Rhabdoviridae	*Indiana vesiculovirus*	5A22	*Vesiculovirus*	*Lyssavirus*	*Lleida bat lyssavirus*	NC_031955	3.25	1.048	-7.16
Picornaviridae	*Cardiovirus A*	4NYZ	*Cardiovirus*	*Parechovirus*	*Parechovirus B*	NC_003976	1.49	0.954	-7.89
Flaviviridae	*Japanese encephalitis virus*	4K6M	*Flavivirus*	*Hepacivirus*	*Equine hepacivirus*	NC_024889	1.41	1.143	-8.11
Reoviridae	*Rotavirus A*	2R7O	*Rotavirus*	*Coltivirus*	*Colorado tick fever coltivirus*	AF133428	0.34	1.134	-9.62
Retroviridae	*Human immunodeficiency virus 1*	5TXL	*Lentivirus*	*Spumavirus*	*Macaque simian foamy virus*	X54482	2.14	1.507	-7.05

Table 4: Homology modelling at intra-family, inter-genus level. Templates are as given in Table 2. Targets are the RdRP (or reverse transcriptase for *Spumavirus*) sequences from the reference genome accession numbers given. RMSD: root mean square deviation in Angstroms between template and model when superposed in MOE. Blue indicates good-quality, red indicates poor-quality, determined by the following thresholds: Φ-Ψ < 2%; QMEAN Z-score > -4.00; RMSD < 2 Å.

Order	Template species	Template PDB	Template family	Target family	Target species	Target reference genome	Φ-Ψ outliers (%)	RMSD (Å)	QMEAN Z-score
Bunyavirales	*La Crosse orthobunyavirus*	5AMQ	*Peribunyaviridae*	*Phenuiviridae*	*Rift Valley fever phlebovirus*	NC_014397	1.98	1.404	-8.99
Mononegavirales	*Indiana vesiculovirus*	5A22	*Rhabdoviridae*	*Bornaviridae*	*Mammalian orthobornavirus 1*	NC_001607	3.53	2.238	-10.06
				Filoviridae	*Zaire ebolavirus*	NC_002549	3.50	1.242	-9.80
					Marburg marburgvirus	NC_001608	2.95	1.460	-10.09
				Paramyxoviridae	*Hendra henipavirus*	NC_001906	3.19	1.333	-9.83
					Measles morbillivirus	NC_001498	2.45	1.309	-9.62
					Mumps orthorubulavirus	NC_002200	3.20	1.494	-9.45

Table 5: Homology modelling at intra-order, inter-family level. Templates are as given in Table 2. Targets are the RdRP (or reverse transcriptase for *Lentivirus*) sequences from the reference genome accession numbers given. RMSD: root mean square deviation in Angstroms between template and model when superposed in MOE. Blue indicates good-quality, red indicates poor-quality, determined by the following thresholds: Φ-Ψ < 2%; QMEAN Z-score > -4.00; RMSD < 2 Å.

Template	**Template PDB**	**Genus**	**Φ-Ψ outliers (%)**	**RMSD (Å)**	**QMEAN Z-score**
Hantaan orthohantavirus	5IZE	*Orthohantavirus*	0.00	0.280	-0.38
Crimean Congo hemorrhagic fever orthonairovirus	3PHX	*Orthonairovirus*	2.37	1.354	-2.88
La Crosse orthobunyavirus	5AMQ	*Orthobunyavirus*	0.45	0.556	-2.64
Indiana vesiculovirus	5A22	*Vesiculovirus*	1.98	0.850	-5.52
Cardiovirus A	4NYZ	*Cardiovirus*	0.86	0.954	-2.15
Rhinovirus A	1XR7	*Enterovirus*	0.22	0.564	-1.10
Lymphocytic choriomeningitis mammarenavirus	3JSB	*Mammarenavirus*	0.00	0.351	-1.43
Japanese encephalitis virus	4K6M	*Flavivirus*	1.21	0.875	-3.94
Hepacivirus C	2YOJ	*Hepacivirus*	1.21	0.701	-2.57
Human picobirnavirus	5I61	*Picobirnavirus*	0.56	0.638	-2.77
Rotavirus A	2R7O	*Rotavirus*	0.20	1.134	-7.09
Human immunodeficiency virus 1	5TXL	*Lentivirus*	1.30	1.507	-2.82

Table 6: Homology modelling the common ancestor for each genus. Templates are as given in Table 2. Targets are the reconstructed ancestral RdRP (or reverse transcriptase for *Lentivirus*) sequences. RMSD: root mean square deviation in Angstroms between template and model when superposed in MOE. Blue indicates good-quality, red indicates poor-quality, determined by the following thresholds: Φ-Ψ < 2%; QMEAN Z-score > -4.00; RMSD < 2 Å.

Level	Φ-Ψ outliers (%)	RMSD (Å)	QMEAN Z-score
Solved structure templates	0.25	N/A	-1.033
Intra-genus, inter-species	1.32 (1.29)	0.839 (0.870)	-5.245 (-5.348)
Intra-family, inter-genus	1.73 (1.72)	1.157 (1.048)	-7.966 (-7.325)
Intra-order, inter-family	2.97	1.497	-9.691
Common ancestor of genus	0.86	0.814	-2.941

Table 7: Mean model (or structure) quality. The top line shows the mean quality scores for the solved structures used. The other lines show the mean quality scores for the models produced at various levels of taxonomic distance between template and target. Blue indicates good-quality, red indicates poor-quality, determined by the following thresholds: Φ-Ψ < 2%; QMEAN Z-score > -4.00; RMSD < 2 Å. Numbers in brackets indicate the revised scores if the model for *Imjin thottimvirus* is moved out of the intra-genus category and into the intra-family category in the light of its subsequent transfer into the new genus *Thottimvirus.*

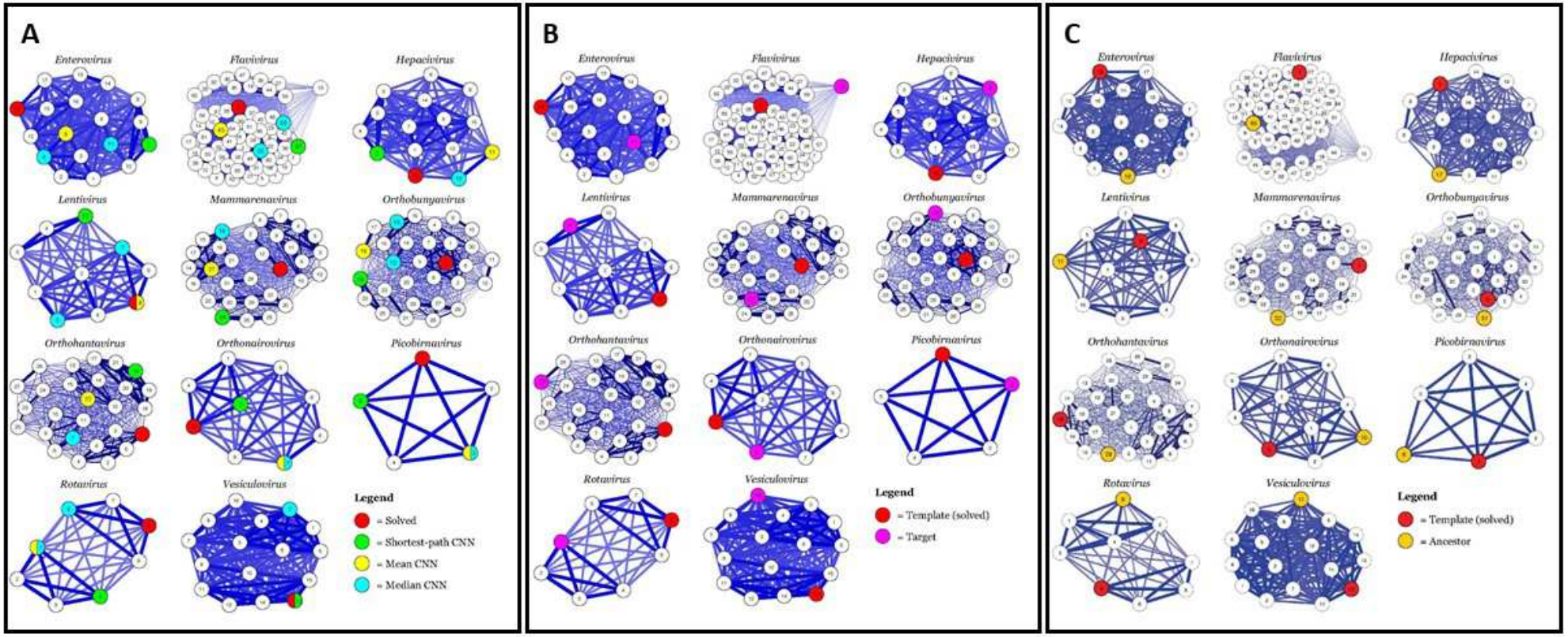


Figure 1: Force-directed graph visualisations of similarity of RdRPs (or reverse transcriptase for *Lentivirus*) within genera

The genetic distance matrix for each alignment was converted into a similarity matrix (Equations 1 and 2). The Fruchterman-Reingold algorithm (500 minimisation iterations) was implemented in R module *qgraph* to produce a force-directed graph. Relative similarity is represented by node proximity, and absolute similarity is proportional to edge thickness. The solved structure and the three types of centroid nearest neighbour (CNN) sequences are highlighted. The species names corresponding to the numbered nodes are listed in the Supplementary Table. *Cardiovirus* has less than four reference sequences and is omitted. A: Location of solved structure and the three CNNs in sequence space (Equations 3-7). Some genera have two median CNNs.

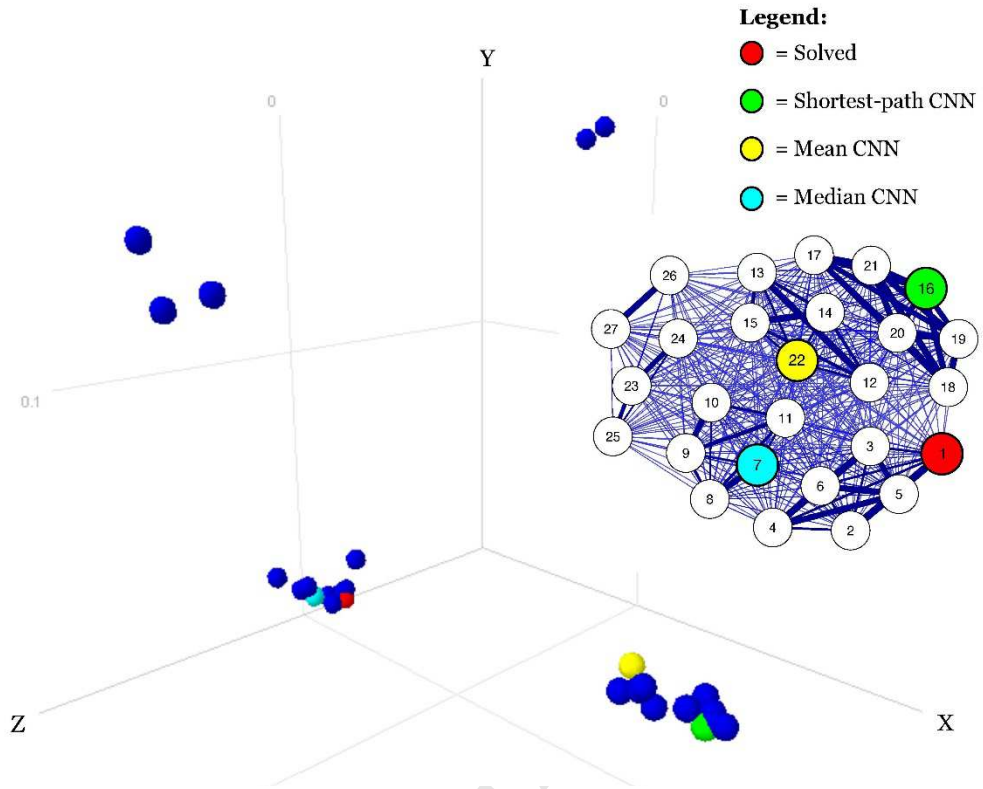


Figure 2: Visualisation of sequence space in two and three dimensions for *Orthohantavirus*

Multi-dimensional scaling on the *Orthohantavirus* similarity matrix was implemented in R module *cmdscale* and viewed in Spotfire Analyst. Inset: the *Orthohantavirus* Fruchterman-Reingold representation from Figure 1. The solved structure and the three types of centroid nearest neighbour (CNN) sequences are highlighted. The species names corresponding to the numbered nodes are listed in the Supplementary Table.

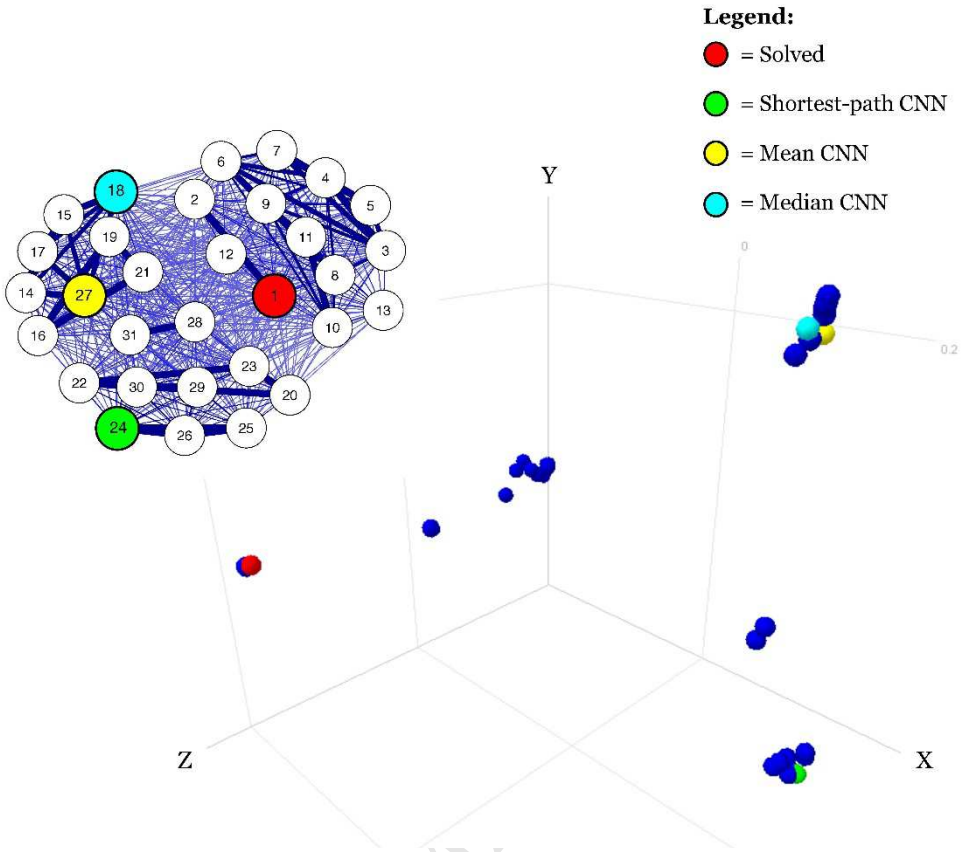


Figure 3: Visualisation of sequence space in two and three dimensions for *Mammarenavirus*

Multi-dimensional scaling on the *Mammarenavirus* similarity matrix was implemented in R module *cmdscale* and viewed in Spotfire Analyst. Inset: the *Mammarenavirus* Fruchterman-Reingold representation from Figure 1. The solved structure and the three types of centroid nearest neighbour (CNN) sequences are highlighted. The species names corresponding to the numbered nodes are listed in the Supplementary Table.

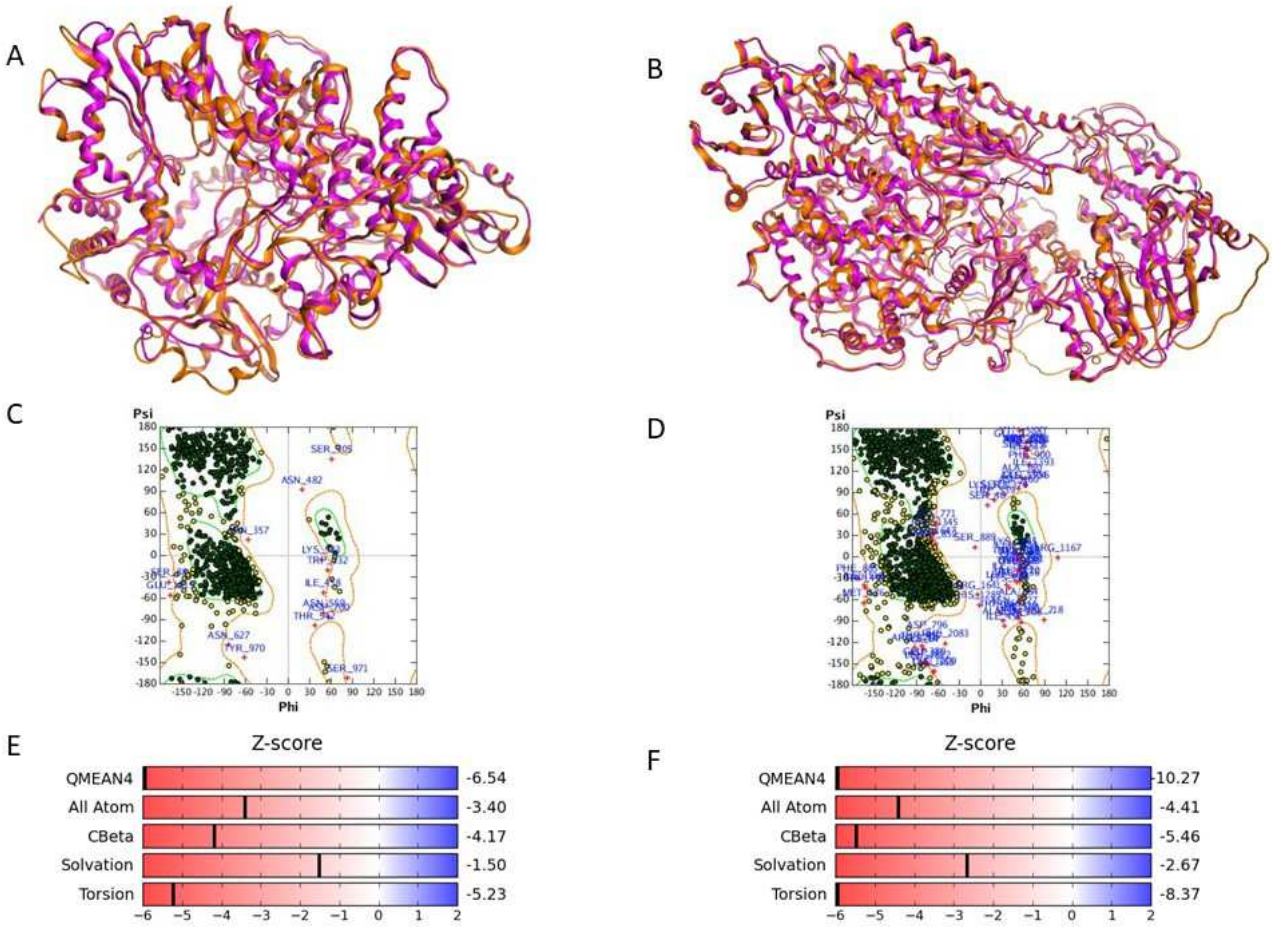


Figure 4: Homology models, Ramachandran (Φ-Ψ) plots and QMEAN Z-scores graphics for the "best" and "worst" intra-genus model

A: Superposition of *Rotavirus I* model (orange) on *Rotavirus A* template 2R7O (pink). B: Superposition of *American bat vesiculovirus* model (orange) on *Indiana vesiculovirus* template 5A22 (pink). C: Ramachandran (Φ-Ψ) plot for *Rotavirus I* model. D: Ramachandran (Φ-Ψ) plot for *American bat vesiculovirus* model. E: QMEAN Z-scores graphic for *Rotavirus I* model. F: QMEAN Z-scores graphic for *American bat vesiculovirus* model. The Φ-Ψ plots (C,D) show Ψ on the y-axis and Φ on the x-axis. Bond angle quality: favoured (green), allowed (yellow), and outliers (red cross, blue text). The Z-score graphics show model quality on a sliding scale: low-quality (red), high-quality (blue). QMEAN4 shows the overall Z-score, "All Atom" shows the average Z-score for all of the atoms in the model, "CBeta" the Z-score for all Cβ carbons, "Solvation" is a measure of how accessible the residues are to solvents, and "Torsion" is a measure of torsion angle for each residue compared to adjacent residues.

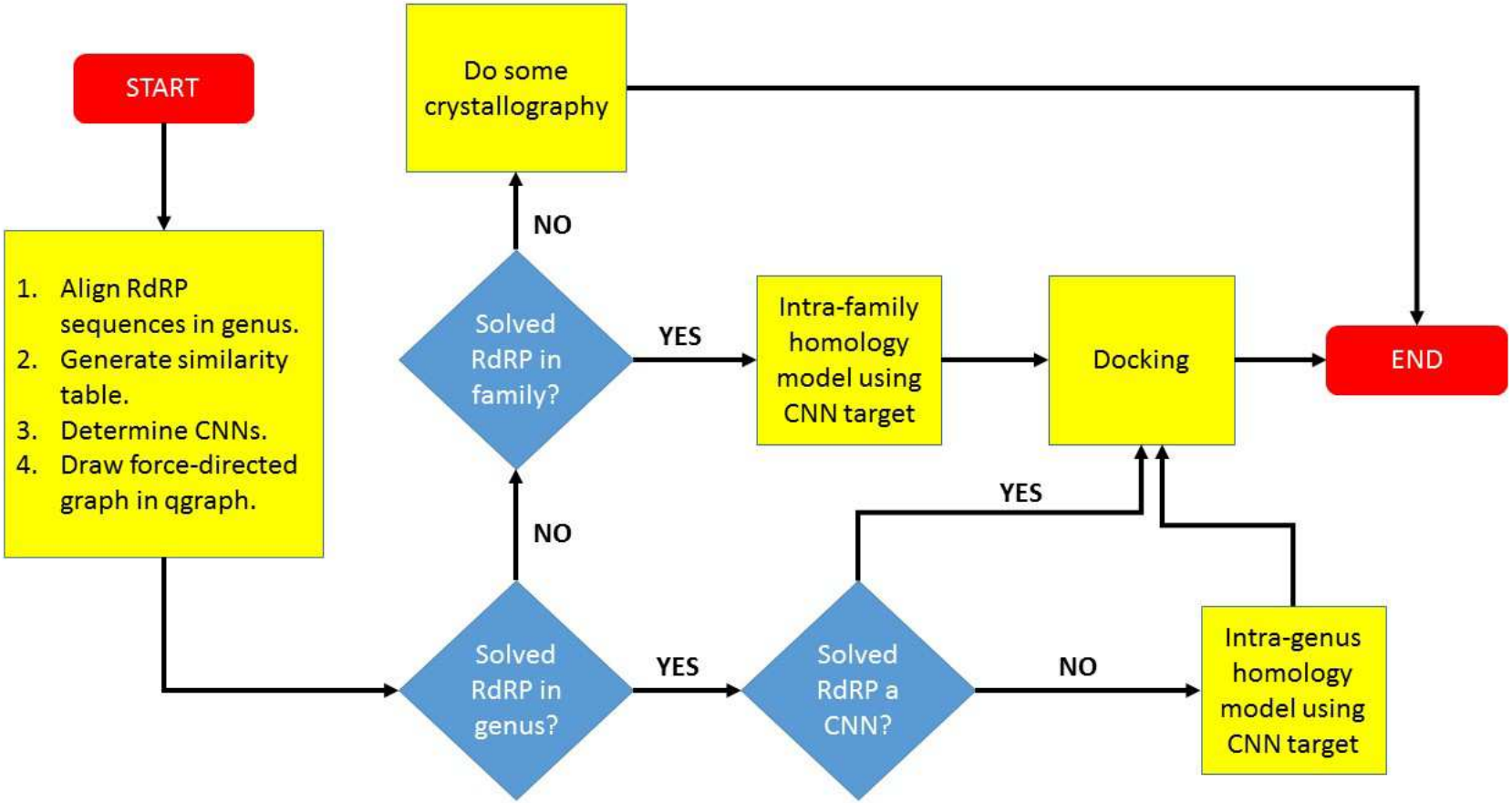


Figure 5: Flowchart of recommended strategy for choice of RdRP for docking experiments.

Where a solved RdRP structure exists in a genus, it should be used. However, if that solved structure is not a CNN, a homology model of a CNN or ancestral sequence should be produced for comparative purposes. Where no solved RdRP structure exists in a genus, a structure from another genus in the same family may be used.

Highlights

1. The first use of force-directed graphs for the visualization of multidimensional protein sequence space in two dimensions
2. Measures of centrality in protein sequence space to identify sequences for production of homology models
3. Homology modelling for RNA-dependent RNA polymerase (RdRP) target-template pairs in different species, genera and families
4. A protocol for the production of optimal RdRP homology models for use in further experiments