ABSTRACT

Motivation: The extraordinary genetic and antigenic variability of RNA viruses is arguably the greatest challenge to the development of broadly effective vaccines. No single viral variant can induce sufficiently broad immunity, and incorporating all known naturally circulating variants into one multivalent vaccine is not feasible. Furthermore, no objective strategies currently exist to select actual viral variants that should be included or excluded in polyvalent vaccines.

Results: To address this problem, we demonstrate a method based on graph theory that quantifies the relative importance of viral variants. We demonstrate our method through application to the envelope glycoprotein gene of a particularly diverse RNA virus of pigs: porcine reproductive and respiratory syndrome virus (PRRSV). Using distance matrices derived from sequence nucleotide difference, amino acid difference and evolutionary distance, we constructed viral networks and used common network statistics to assign each sequence an objective ranking of relative 'importance'.

To validate our approach, we use an independent published algorithm to score our top-ranked wild-type variants for coverage of putative T-cell epitopes across the 9383 sequences in our dataset. Top-ranked viruses achieve significantly higher coverage than low-ranked viruses, and top-ranked viruses achieve nearly equal coverage as a synthetic mosaic protein constructed in silico from the same set of 9383 sequences.

Conclusion: Our approach relies on the network structure of PRRSV but applies to any diverse RNA virus because it identifies subsets of viral variants that are most important to overall viral diversity. We suggest that this method, through the objective quantification of variant importance, provides criteria for choosing viral variants for further characterization, diagnostics, surveillance and ultimately polyvalent vaccine development.

Contact: tgoldberg@vetmed.wisc.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on December 13, 2011; revised on March 10, 2012; accepted on April 4, 2012

1 INTRODUCTION

Of all infectious threats to global health, viruses with RNA genomes have proven to be the most difficult to control (Cleaveland et al., 2001; Jones et al., 2008). Viruses such as HIV, influenza virus, hepatitis C virus and Dengue virus have defied control in part as a result of their extraordinary genetic and antigenic variation (Katze et al., 2008) and their consequent ability to evolve rapidly (Belshaw et al., 2008; Domingo, 2002; Drake and Holland, 1999; Holmes, 2009). HIV, for example, can evolve into diverse strains within an individual patient over the course of days or weeks, confounding the development of vaccines and antivirals and frustrating diagnostics and surveillance (Crandall, 1999; Gaschen et al., 2002). Thus, the efficacy of even the most immunogenic vaccines is questionable in a field setting where multiple variants may circulate simultaneously and where novel variants are continuously evolving. Vaccine studies almost universally acknowledge this challenge but have yet to develop objective criteria selecting vaccine candidates that adequately, yet parsimoniously, capture overall viral diversity (Barouch, 2008; Korbet et al., 2009).

In an effort to develop vaccines that maximize the representation of antigenic features present in diverse viral populations, a series of computational strategies have been proposed. Driven largely by HIV vaccine research, the approaches have included concatenating commonly recognized T-cell epitopes (Palker et al., 1989), creating pseudoprotein strings of T-cell epitopes (De Groot et al., 2005) and generating consensus overlapping peptide sets from proteins (Thomson et al., 2005). Evolutionary approaches such as the use of consensus sequences (Gao et al., 2004, 2005; Gaschen et al., 2002), and the most recent common ancestor (MRCA) of viral populations, have also been proposed with the assumption that these approaches capture viral diversity (Gaschen et al., 2002). Unfortunately, experimental studies in animal models using these strategies have documented overwhelming humoral immune responses (Doria-Rose et al., 2005; Gao et al., 2005).
The most widely studied approach uses a genetic algorithm to generate, select and recombine (in silico) potential T-cell epitopes into full-length ‘mosaic’ protein sequences that can provide greater coverage of global viral variants than could any single wild-type protein (Fischer et al., 2007). This mosaic protein approach was able to achieve between 74% and 87% coverage of HIV-1 Gag sequences, compared with only between 37% and 67% using a single natural Gag protein (Fischer et al., 2007). Subsequent experimental studies have demonstrated that these computationally designed vaccines augment the breadth and depth of antigen-specific T-cell responses as compared with consensus or natural sequence HIV-1 antigens (Barouch et al., 2010; Santra et al., 2010). However, this approach yields synthetic peptides with unknown biological properties that may not be able to be incorporated into a recombinant live vaccine.

Given the need for an objective strategy to select wild-type viral variants that might be adapted for polyvalent vaccines, we propose a method based on network analysis. In recent years, interest in the analysis and modeling of networks has surged outside the traditional fields of social science (Wasserman and Faust, 1994; Watts, 2004), mathematics and computer science (Albert and Barabasi, 2002; Butts, 2009). Network statistics have proven to be powerful tools for studying diverse phenomena such as social interactions, connections among neurons and the structure of the Internet, and for optimizing solutions to engineering problems (Albert and Barabasi, 2002; Watts, 2004). For our purpose, it has long been recognized that biological molecules interact directly and indirectly (Yamada and Bork, 2009), and network statistics have proven fruitful in drug discovery studies (Arzum et al., 2010; Zhao and Li, 2010), human disease classification (Barabasi et al., 2011; Loscalzo et al., 2007) and functional analyses of gene regulatory networks (Yu et al., 2007). As such, it is possible to construct viral networks that explicitly define the positional importance of individual viral variants in the network based on relationships between biomolecules (e.g. Allesina and Pascual, 2008; Snape et al., 2010; Yamada and Bork, 2009).

In this article, we use a highly diverse and particularly problematic RNA virus of pigs as a case study: porcine reproductive and respiratory syndrome virus (PRRSV). This exercise provides ‘proof of concept’ as well as a novel strategy for the control of this virus of critical importance to agriculture. We present a detailed topological analysis network that considers the importance of every individual viral variant \( i \) by considering every possible interaction (direct and indirect) between variant \( i \) and \( j \). We use network indices, from local to global, including node degree, a series of centrality indices and PageRank, the algorithm at the heart of Google. We then examine how these indices relate to each other and discuss how to choose the most appropriate metric for selecting viral variants for study and to help guide the development of polyvalent vaccines, diagnostic testing strategies and epidemiological surveillance. Our study makes minimal initial assumptions about evolutionary processes or viral biology and can complement other approaches such as the mosaic protein approach (Fischer et al., 2007) to polyvalent vaccine design.

2 METHODS

2.1 Porcine reproductive and respiratory syndrome virus

PRRSV is the causative agent of porcine reproductive and respiratory syndrome (PRRS), an economically damaging disease of domestic swine (Tian et al., 2007; Zimmerman et al., 1997). PRRSV is in the order Nidovirales, family Arteriviridae, along with equine arteritis virus, lactate dehydrogenase elevating virus and simian hemorrhagic fever virus (Conzelmann et al., 1993; Meulenberg et al., 1993). The virus has a short (~15 kb), single-stranded, positive-sense RNA genome that encodes at least nine open reading frames (ORFs; Conzelmann et al., 1993; Meulenberg et al., 1993). Despite significant research, including a focus on the genetic diversity of the virus, it has remained difficult to control (Shi et al., 2003a,b). Vaccination in the field has variable efficacy and has failed to reduce transmission or to reduce clinical signs (see Thanawongnuwech and Suradhat, 2010 and references therein).

One likely cause for the difficulty in controlling PRRSV is its genetic and antigenic variability. Type 1 PRRSV (or European origin) differs at ~50% of nucleic acid positions from Type 2 PRRSV (of North American origin), implying a high degree of divergence between continents (Forberg et al., 2001; 2002; Muttera et al., 2010). Further, between viral types, PRRSV exists as a spectrum of genotypes, displaying considerable heterogeneity with very little geographic population substructure (Shi et al., 2010a,b; Stadejek et al., 2006). Additionally, PRRSV varies between farms, herds, among animals within herds and even within individual animals (Goldberg et al., 2000, 2003). Extensive nucleotide sequence data are available for the viral ORF5 gene, which is used commonly for phylogenetic analyses and source tracking (e.g. Shi et al., 2003a,b). Consequently, ORF5 diversity acts as a marker of PRRSV variability.

2.2 Creation of a core PRRSV sequence dataset

All PRRSV ORF5 sequences from isolates in Genbank, the National Center for Biotechnology Information’s online sequence repository (http://www.ncbi.nlm.nih.gov/Genbank/), and the now defunct PRRSV Database were downloaded in December 2010 (see Supplementary Material for sequence files). These sequences comprised 25,265 worldwide samples, 3 live-attenuated vaccine strains (MLV, ATP and PrimePac) and one laboratory attenuated strain (Abst-1). Most viruses were from the United States of America, but viruses from Canada, Mexico, China, Korea, Japan, Thailand, Austria, Denmark, Italy and Poland were also present.

As a first ‘cleaning’ step to ensure we only included wild-type variants, we removed sequences that were described with any of the following terms: patent, vaccine, attenuated, attenuation and clone. The remaining sequences were then aligned using default settings in MUSCLE v3.6.6 (Edgar, 2004), with manual correction in Mesquite (Maddison and Maddison, 2011). At the same time, we removed all type 1 variants, thus restricting our analyses to type 2 PRRSV. The remaining aligned sequences underwent a redundancy analysis within the computer program Mothur (Schloss et al., 2009), and identical sequences were removed. In addition, to avoid bias in subsequent network and phylogenetic analyses (Lemmon et al., 2009), we removed sequences with any nucleotide base ambiguities. The next filtering step was to remove poor quality data using three criteria: first, sequences were removed if they did not have a start and a stop codon; second, sequences were removed if >1% of the gene sequence was missing; and third, a sequence was removed if it produced an insertion or deletion in the alignment that was biologically implausible (e.g. causing a frame shift and introducing premature stop codons) or occurred in fewer than 0.005% of sequences. The alignment was then screened for evidence of recombination using 3Seq (Boni et al., 2010), with all putative recombinants subject to secondary screening and validation using RDP3 (Martin et al., 2010). If identified by three or more methods within the software, sequences were removed from further analysis.

For reference, even though some were eliminated in our filtering process, we included several common reference strains: VR2332, ResPrrsv, China ‘atypical’, IA-142, MN-184, PrimePac and ATP viral variants. This process resulted in a set of 9383 non-identical ORF5 sequences that represent the full extent of known genetic diversity of wild-type type 2 PRRSV.

2.3 Construction of PRRSV networks

We constructed three undirected, unweighted networks using adjacency matrices derived from (i) nucleotide sequence differences, (ii) amino acid differences and (iii) network centralities.
sequence differences and (iii) evolutionary distances. We generated the nucleotide difference adjacency matrix using PAUP* 4.0 (Swoford, 2003), with cells containing values representing uncorrected numbers of nucleotide differences between pairs of aligned sequences. Similarly, we generated a second adjacency matrix by calculating the uncorrected number of changes between amino acid sequences using program R (Team, 2010) with the APE (Paradis et al., 2004) and SeqinR packages (Charif and Lobry, 2007). Our third adjacency matrix represented evolutionary distances. Values in cells were patristic distances (distances along the branches of a phylogenetic tree) calculated by PAUP* 4.0 (Swoford, 2003); we built the phylogeny used to calculate these distances using RAxML 7.2.8 (Stamatakis, 2006, 2008) maximum likelihood analyses on the CIPRES Science Gateway (Miller et al., 2010). We implemented a General-Time-Reversible nucleotide substitution model with four categories of gamma distributed rate heterogeneity and a proportion of invariant sites (GTri+I+Γ) model of molecular evolution.

The three matrices represent different types of biological relationships among viral sequences. The first makes the fewest assumption about the structure or function of each variant, the second acknowledges that immunological cross-protection and virulence operate at the level of proteins, and the third acknowledges that evolutionary history might be an important constraint on viral immunobiology. Although these matrices contain quantitative information on relationships between viral variants, to simplify theory and measurement, we dichotomized each using the snA package in R (Butts, 2008). This approach is appropriate given that the relationship under study, in this instance genetic distance, is stable and the values taken across pairs are constrained (Butts, 2009).

2.4 Topological indices used to describe positional importance

To rank sequences, we used a range of indices that are dependent on the characteristics of the focal node and also included information about the overall topology of the viral network.

The most fundamental metric, the degree (or connectivity; k) of a sequence, describes the number of links a sequence makes with other sequences. Those sequences with high degrees are those that are connected highly with all other sequences. This index is useful for identifying viral variants that have the most direct connections to other variants. Other, secondary, clustering techniques further detail the relative importance of each sequence to ‘sequence diversity space’ and include closeness; betweenness; eigenvector centrality; and PageRank.

Closeness provides a measure that describes the position, in terms of distance, from a focal sequence to all other sequences and is measured as:

$$\text{CC}(v) = \frac{1}{\sum_{u \neq v} d(v, u)} \quad (1)$$

Intuitively, closeness provides an index of the extent to which a given sequence has short paths to all other sequences. This is a measure of whether a sequence is in the ‘middle’ of our defined sequence space; smaller values can be interpreted to indicate sequences of greater importance (i.e. the removal of such a sequence will affect a majority of other sequences).

Betweenness, first introduced by Freeman (1977), is a measure that describes the centrality of a sequence, provided as a frequency with which a sequence is located on the shortest path between all other sequences. First, let \( \delta_{ij}(v) \) denote the pairwise dependency or the fraction of shortest paths between i and j that pass through v. Following this, the betweenness of virus \( v \) is defined as:

$$B(v) = \sum_{i,j \neq v, i \neq j} \delta_{ij}(v) \quad (2)$$

Conceptually, high-betweenness sequences lie on a large number of non-redundant shortest paths between other vertices; they may be thought of as ‘bridges’. This metric has been used to assess node importance in other biological networks (Yamada and Bozk, 2009), to describe sexual contact networks (Liljeros et al., 2001) and to describe the structure of the worldwide air transportation network (Guimerà et al., 2005).

Eigenvector centrality, as described by Bonacich (1972), determines node scores that correspond to the leading eigenvector of the sequence adjacency matrix. These scores arise from a reciprocal process in which the centrality of each sequence is proportional to the sum of the centralities of those sequences to which it is connected. The defining equation is as follows:

$$x_i = \lambda x_i \quad (3)$$

where \( A \) is the adjacency matrix of the graph, \( \lambda \) is the constant (the eigenvalue) and \( w \) is the eigenvector. In general, sequences that have high eigenvector centralities are those that are connected to many other sequences that are, in turn, connected to many others. This metric has been widely used to identify key network components, including connectivity within the human brain (Lohmann et al., 2010). PageRank, in the context in which we use it, rates viral variants as important if they receive links from other viruses that are in turn also rated as important. Each virus i is assigned an importance, and each link \( a_{ij} \) (existing virus i to enter virus j) carries an equal fraction of the importance value: the importance of the virus is then the sum of the importance assigned to its incoming connections. This recursive process is solved by building a matrix \( S \) in which each element represents the fraction of importance assigned to a link: subsequently, importance is solved by computing the eigenvector associated with the dominant eigenvalue of the matrix (Bryan and Leise, 2006).

Finally, we analyze the community structure of the networks by computing the modularity \( Q \) given by (Clauset et al., 2004):

$$Q = \sum_{i=1}^{m} \left( e_{ii} - a_{ii} \right) \quad (4)$$

where \( m \) is the number of modules inside the network, \( e_{ii} \) is the fraction of links in the network connecting nodes of the same community i and \( a_{ii} \) is the fraction of links that have one or two ends inside community i. The larger the fraction of links inside each community, the higher the value of \( Q \). In this way, modularity \( Q \) can be taken as a reference parameter to find the optimal community divisions based on the topology of the network. These algorithms have been used extensively to address interaction patterns within varied systems, including but not limited to American football (Girvan and Newman, 2002), and plant-pollinator dynamics (Olesen et al., 2007).

The critical concept to note in the above analytical approach is that the position a sequence takes within ‘sequence diversity space’ is determined by its similarity to other sequences and the similarity values of the other sequences to which it is connected. In using these algorithms, we account for the underlying structure of the genetic relationships between sequences and can objectively rank each sequence and its contribution to PRRSV genetic diversity.

2.5 Relationship between indices

We calculated the indices described above for every individual viral variant in each PRRSV network. We ranked the importance of all variants according to their values of a particular index. Thus, in total there were five rankings corresponding to five different indices (i.e. k, CC, B, w and PR) for each of three networks. We also examined whether or not different indices provided the same information about the relative importance of viral variants. First, to show the relationship between indices, we conducted hierarchical clustering on the rankings of the viral variants by the values of the five indices: first, we standardized the values within indices and then used similarity measures appropriate for interval data. Second, we used Kendall rank correlations between all pairs of indices; and third, we applied a Kendall Concordance test to determine whether the observed ranks were significantly different from those expected by chance when independently ranking 9383 nodes by five indices.

2.6 Design and evaluation of T-cell vaccine candidates

We compared the putative epitope coverage of known PRRSV variants by generating vaccine ‘mosaic’ proteins using a promising design algorithm.
from the HIV literature (Fischer et al., 2007). To assess whether high-ranking variants achieved better epitope coverage than low-ranking variants, we used the Vaccine Epitope Coverage Assessment online tool (Epipoccer; Thurmord et al., 2008) to compute how well the top- and bottom-ranked proteins from each network covered putative epitopes across the full dataset of 9383 wild-type PRRSV sequences. To further explore the association between network ranking and epitope coverage, we ranked sequences according to each centrality metric, selected every 20th ranked sequence and calculated its epitope coverage. We quantified the strength of association using MIC (Reshef et al., 2011) and Pearson’s correlation coefficient ($\rho$). To compare our method to the mosaic protein approach (Thurmond et al., 2008), we used the Mosaic Vaccine Designer online tool to generate in silico a single mosaic protein from the 9383 wild-type sequences. We then used Epipoccer to measure the coverage of this single mosaic across the 9383 wild-type sequences.

3 RESULTS

The viral network derived from nucleotide differences included 9383 variants and had 88,040,689 potential links of which 11,040,560 were realized, resulting in a connectance of 0.125. The amino acid-derived network included 7889 variants and had 88,040,689 potential links of which 11,040,560 were realized, resulting in a connectance of 0.125. The evolutionary distance network included 36 variants and had 11,040,560 potential links of which 11,040,560 were realized, resulting in a connectance of 0.125. Measured in this way, connectance is the average fraction of viral variants to which an individual variant in the network is connected based on similarity. Twenty-two viral variants were ranked in the top 100 sequences in all three networks, and 51 variants were ranked in the bottom 100 sequences in all three networks.

All three viral networks displayed cumulative degree distributions that were different from what would be expected if the link distribution were random. Each network had data consistent with an exponential degree distribution $P(k) \sim \exp(-\gamma k)$; nucleotide $AICc = -8449.9$; evolutionary distance $AICc = -7637.8$; amino acid $AICc = -6191.2$. The data were not well represented by the power law $P(k) \sim k^{-\gamma}$: nucleotide $AICc = -1246.5$; evolutionary distance $AICc = -1417.4$; amino acid $AICc = -816.1$. No other models were found to be of similar quality. Nevertheless, the rank orderings of variants were statistically significantly correlated between indices (i.e., a variant ranked highly by degree ($k$) tended also to receive a high ranking as measured by eigenvector centrality, Table 2). This was further validated using Kendall’s $W$, which showed that the ranking provided by each index is significantly similar across each viral network: nucleotide differences ($W = 0.92$, $x^2 = 38421$, $P < 0.01$); amino acid differences ($W = 0.85$, $x^2 = 33495$, $P < 0.01$) and evolutionary distance ($W = 0.798$, $x^2 = 37447$, $P < 0.01$).

Our ranking approach indentified a single wild-type sequence that achieved coverage of putative epitopes essentially equal to that of a ‘mosaic’ protein generated using all 9383 sequences (Fig. 2). Our top-ranked variant (DQ475317) within the nucleotide network achieved 48.5% coverage, whereas the mosaic protein achieved 49%. Top-ranking variants consistently covered more putative epitopes than those that were ranked poorly; the top-ranking amino acid variant achieved 40% coverage; and the most important variant in the evolutionary distance network achieved 41% coverage (Fig. 2). The 22 variants ranked in the top 100 of all three networks achieved, on average, 43% ($\pm 1.7\%$ SD) coverage of putative epitopes derived from known PRRSV ORF5

### Table 1. Network ranking of select porcine reproductive and respiratory syndrome virus reference sequences and the top five ranked sequences out of 9383

<table>
<thead>
<tr>
<th>Reference variants</th>
<th>Nucleotide network</th>
<th>Amino acid network</th>
<th>Evolutionary distance network</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResP9RS (AP066183)</td>
<td>3915 2857 4478 4279 4875 5666 4627 5447 3814 5624 5064 2807 1520 4441 3584</td>
<td>3915 2857 4478 4279 4875 5666 4627 5447 3814 5624 5064 2807 1520 4441 3584</td>
<td>3915 2857 4478 4279 4875 5666 4627 5447 3814 5624 5064 2807 1520 4441 3584</td>
</tr>
<tr>
<td>JA-142 (AY424271)</td>
<td>595 571 530 780 554 31 39 91 79 31 445 511 743 485 534</td>
<td>595 571 530 780 554 31 39 91 79 31 445 511 743 485 534</td>
<td>595 571 530 780 554 31 39 91 79 31 445 511 743 485 534</td>
</tr>
<tr>
<td>MX-164 (EF487899)</td>
<td>3605 4619 1881 6497 1367 3352 4303 4484 3080 3097 3107 1259 997 6327 1459</td>
<td>3605 4619 1881 6497 1367 3352 4303 4484 3080 3097 3107 1259 997 6327 1459</td>
<td>3605 4619 1881 6497 1367 3352 4303 4484 3080 3097 3107 1259 997 6327 1459</td>
</tr>
<tr>
<td>PrimEPA (A0666384)</td>
<td>586 489 533 701 583 202 221 206 468 133 136 46 482 94</td>
<td>586 489 533 701 583 202 221 206 468 133 136 46 482 94</td>
<td>586 489 533 701 583 202 221 206 468 133 136 46 482 94</td>
</tr>
<tr>
<td>China ‘atypical’ (EF112446)</td>
<td>2732 1909 3387 2467 4248 5352 4120 3153 3295 5317 1448 2354 2866 1447 1911</td>
<td>2732 1909 3387 2467 4248 5352 4120 3153 3295 5317 1448 2354 2866 1447 1911</td>
<td>2732 1909 3387 2467 4248 5352 4120 3153 3295 5317 1448 2354 2866 1447 1911</td>
</tr>
</tbody>
</table>

Top five ranked variants:

- **DQ475317**
- **DQ477864**
- **DQ477862**
- **AB175723**

Ranking statistics include degree ($k$), closeness centrality (CC), betweenness (B), eigenvector centrality (w) and PageRank (PR). Variants are referred to using common name or GenBank accession numbers. Rankings are on a scale of 1 to 9383. The top five sequences are a selection of those variants that were found ranked in the top 100 in all three networks using degree ($k$) as the ranking metric.
variants. By contrast, the 51 variants ranked in the bottom 100 of all three networks achieved, on average, only 26% (±3.5% SD). This difference was statistically significant \((t = 28.20; 69 \text{ degrees of freedom}; P < 0.0001)\). There was a significant \((P < 0.0001)\) relationship between network ranking and putative epitope coverage when variants were ranked by degree (Fig. 3; \(\rho = -0.76; \text{MIC} = 0.63\)); closeness (\(\rho = -0.78; \text{MIC} = 0.61\)); betweenness (\(\rho = -0.66; \text{MIC} = 0.52\)); PageRank (\(\rho = -0.65; \text{MIC} = 0.47\)) and eigenvector centrality (\(\rho = -0.59; \text{MIC} = 0.53\)).

4 DISCUSSION

Our results demonstrate that a dataset of 9383 sequences could be ranked according to three network centrality metrics. Using interaction networks derived from sequence nucleotide difference, amino acid difference and evolutionary distance, we were able to score individual PRRSV sequences based on their relative importance to the viral network. This yielded 22 sequences that were represented in the top 100 sequences in all matrices and provided objective criteria for selecting viral variants for polyvalent vaccine design. We identified a single wild-type sequence that covered 48% of putative epitopes derived from the known diversity of PRRSV. Furthermore, we document a relationship between our ranked variants and putative epitope coverage. In summary, we have demonstrated that it is possible to use network statistics to rank viruses with respect to their overall ‘importance’ within a network and that the resulting rankings can help guide the selection of viruses for inclusion in vaccines.

The key insight provided by our work is that it is possible to construct de novo a virus network from sequence data and objectively quantify the positional importance of viral variants within such a network. For vaccine development, it is clear that choosing a single variant as a type strain is problematic, and vaccination with all known viral variants is not technically feasible. Consequently, there is a need for techniques that ‘filter’ highly diverse genomic datasets and select representative sequences for further study. We suggest that the calculation of network indices that describe the relative importance of viral variants may inform decisions about type strain selection and polyvalent vaccine development. Although this approach is flexible and can be implemented using small and less diverse genomic data, more traditional approaches may be more appropriate in these situations. Of the indices we measured, eigenvector centrality provided a robust prediction of variant importance. This metric provides a measure of a viral variant’s global position, rather than its sheer number of connections (i.e. \(k\)), and describes sequences that are highly connected and fall within densely populated substructures of the viral network. More generally, our case study of the PRRSV system serves as an illustration of the potential of these techniques in the objective selection of important viral variants for additional study.

We demonstrate that it is possible to select only highly ranked sequences and achieve equal or better coverage of putative T-cell epitopes than using low-ranked sequences and that top-ranked sequences achieve nearly as good T-cell epitope coverage as artificial mosaic proteins derived computationally from the full set or sequences. The latter conclusion is important because mosaic...
These viruses, if they can be found in reference collections, would make excellent candidates for subsequent biological studies, with the mosaic protein, were compared against all nine-mers generated from 9383 unique PRRSV variants.

Our approach is likely to be useful given the extraordinarily large amounts of data currently available as a result of next-generation sequencing technologies (Holmes and Grenfell, 2009). Although existing computational methods to account for viral diversity are powerful when investigating small numbers of sequences, the size of potential tree space and computational complexity increases faster than exponentially with the number of sequences (Holmes and Grenfell, 2009; Rambaut et al., 2008). This forces use of biology-independent filters (e.g. limiting analysis to geographic regions, considering ease-of-access to samples or using well-characterized independent filters (e.g. limiting analysis to geographic regions, considering ease-of-access to samples or using well-characterized independent filters (e.g. limiting analysis to geographic regions, considering ease-of-access to samples or using well-characterized independent filters (e.g. limiting analysis to geographic regions, considering ease-of-access to samples or using well-characterized independent filters (e.g. limiting analysis to geographic regions, considering ease-of-access to samples or using well-characterized independent filters (e.g. limiting analysis to geographic regions, considering ease-of-access to samples or using well-characterized independent filters (e.g. limiting analysis to geographic regions, considering ease-of-access to samples or using well-characterized independent filters (e.g. limiting analysis to geographic regions, considering ease-of-access to samples or using well-characterized independent filters (e.g. limiting analysis to geographic regions, considering ease-of-access to samples or using well-characterized independent filters). Although biases may be introduced to the genetic networks due to limitations in source data, our approach is fully adaptable to sample sets with different or no biases, such as geographically restricted populations, or to virulent populations of viruses. Notably, our network analyses were conducted and completed in 72 h on a standard 2 GHz desktop computer with 8 GB of RAM. This suggests that our framework is not prohibitive from a computational perspective.

We suggest that network centrality measures, in differentiating viral variants according to how influential they are, can inform the selection of type strains for vaccine design. The degree, \( k \), of a node is the first indication of its centrality; intuitively, a well-connected sequence with a high degree will be traversed by a proportionally larger number of short paths than those sequences with low degrees. However, it is important to distinguish this metric as a local index; it does not take in to account the importance of the neighboring nodes. This suggests that a variant, although well connected, may fall within a relatively sparse area of the network and may not account for a large amount of genetic information. To overcome this restriction, the other centrality indices quantify positional importance using global measures accounting for genetic distance (i.e. the number of edges on the shortest path connecting \( i \) and \( j \)). Eigenvector centrality is particularly useful; previous research has suggested that nodes characterized by higher eigenvector centrality play a more influential role in the structure and dynamics of the networks in which they are nested (Allesina and Pascual, 2008, 2009). In our viral networks, those variants that have high eigenvector centrality scores are those that contribute disproportionately to global viral diversity, acting as core variants that are likely to prove useful in the development of vaccines, diagnostic tests and epidemiological surveillance strategies.

Another potential metric that provides insight into variant selection is the betweenness of a sequence. One biological interpretation of this index is that it quantifies the probability that...
node $i$ represents an intermediate step in the evolution between one sequence to another. In Table 2, we demonstrated a positive correlation between sequence degree and sequence betweenness, confirming the intuitive idea that sequences of high degree are those with high betweenness: the larger the number of neighbors a particular sequence has, the greater the probability that the sequence will fall on a shortest path between other sequences. Deviations from this correlation indicate ‘gatekeeper’ sequences, i.e. sequences that represent unique pathways to substructures within the network. Thus, using this metric, we are able to quantify a variant’s importance in two ways: first, a variant that has a high betweenness score provides a measure of the amount of genetic information which the variant controls; and second, if a variant falls outside of the correlation we document, it represents a unique sequence that acts as a bridge to distinct modules within the viral network.

Our rankings of PRRSV reference variants reveal that many studies of PRRSV biology to date have focused on reference strains that are not central to the diversity of the virus. Typically, such reference strains are selected because of ease-of-access, publication history and evaluation of whether the variant represents a particular geographic location or biological property (e.g. virulence and immunogenicity). Our analysis shows that selection of such strains may limit the generality of subsequent conclusions. Additionally, current phylogenetic methods that attempt to classify type strains (Shi et al., 2010a,b), although analytically rigorous, carry a series of assumptions that introduce biases into variant selection. This is particularly true in PRRSV, where incomplete geographic sampling of assumptions that introduce biases into variant selection. We believe our approach is best suited for PRRSV and other viruses with a large amount of genetic diversity and little phylogenetic substructure. In the case of viruses that sort into clear monophyletic clades, it may be impossible to select one variant that represents diversity across clades. In this situation, we suggest a ranking system based on specific scenarios, such as geographically restricted populations, virulent populations of viruses or, as in our case study, the selection of known subtypes. An alternate solution is to select sequences from within defined network groups using modularity-based clustering heuristics (see Newman, 2012 and references therein).

One strength of our approach is that it makes minimal assumptions about viral biology in assigning the rankings of individual sequences. Underlying matrices of amino acid similarity, for example, do not assign immunological importance to specific protein regions. However, our approach is very flexible for incorporating biological information as it is generated. For example, our approach could easily be applied to specific protein regions that future studies identify as being immunologically important. A weakness, however, is that the representational framework is restrictive given the dyadic nature of interactions; if our goal is to develop an effective vaccine, this process may omit components of what is a complex system of interactions. Indeed, research in the HIV field has revealed that induction of protective immunity is a function of robust mucosal immunity, high avidity and polyfunctional T-cells, broad neutralizing antibodies and optimized vaccine delivery methods (Wijesundara et al., 2011). This is in addition to the difficulties associated with viral antigenic diversity (Fischer et al., 2007).

However, decades of research have been unsuccessful in the development of broadly effective vaccines for numerous highly variable pathogens. Thus, our approach to representing variability as a network, and ranking variants for further study, may prove to be a judicious strategy that aids in the development of efficacious vaccines and targeted control.

Important tasks for future studies include: (i) improving our knowledge of the biological meanings of different centrality indices; (ii) extending and improving the ranking system through comparative analyses of similar viruses; and (iii) testing whether the high-ranking individual variants are better in inducing cross-protective immunity than those with lower rankings in vivo. Furthermore, a potential and as yet unrealized application of our approach is the targeted design of vaccines that modify virulence. Our case study incorporated all sequence variation, but it is possible to select for analysis only sequences of high virulence. Vaccines rarely provide full protection from disease, and those vaccines that are ineffective may unintentionally select for increased pathogen virulence (Gandon et al., 2001). We posit that, given the high likelihood that any vaccine against a highly variable RNA virus such as PRRSV would offer less-than-perfect protection, an alternative goal of vaccination could be to drive viral evolution intentionally toward benignness. Our ranking method could provide rational criteria for selecting among highly virulent types to include in vaccines that may, over time, drive the population of viruses toward lower virulence (Ewald, 2004).

ACKNOWLEDGEMENTS

Funding: PRRS CAP, United States Department of Agriculture NIFA Award 2008-55620-19132.

Conflict of Interest: none declared.

REFERENCES


Forsberg, R. et al. (2005) The genetic diversity of European type PRRSV is similar to that of the North American type but is geographically skewed within Europe. Virology, 340, 36–47.


