Results

Sequence variability in viral proteins is slight — most PDB structures have less than 11 related sequences different at 2% of aligned amino acids, 7 or less at 5% difference, and 4 or less at 10% difference (Table 1). In Dengue, West Nile, Hepatitis C, Japanese Encephalitis, and Hemaglutinin proteins where more than 3 PDB structures are present at the 5% difference, further analysis was done to establish a relationship between amino acid sequence variation and their root mean square fluctuations (Dengue and Japanese Encephalitis proteins had practically the same resulting sequences, thus only Japanese Encephalitis was used for this analysis). The correlation here is at most about 32% (Fig 1).

Methods

Firstly, each protein structure of interest is broken down into chains to assess which of them are suitable for BLAST. This analysis is based on the similarity a chain conveys in regards to other chains in the structure. In the case where two or more chains are identical, only one chain is used for further analysis. Similarly, when two chains are "identical" but one has an extra tail of amino acids, the longer version of the chain is used. When chains differ entirely, both chains are used. Chains and structures used in this analysis are presented in Table 1.

Filtered chains are ran through BLAST using `blast.pdb()` function in R. The results are presented in a Hit Table format from smallest to largest e-value. Three variables from the Hit Table are considered for this analysis: sequence identity, alignment length, and bit score. Sequence identity indicates the similarity between sequences at aligned residue positions. Alignment length is the number of residues

that have been aligned. Bit score is an indicator of a successful alignment with the higher scores suggesting better alignment. Bit score is also normalized allowing for different alignments to be compared regardless of what alignment matrix was used for the alignment.

Next, found sequences are filtered based on the criteria of ≥40% sequence identity and ≥90% alignment length; however, some alignments (West Nile Virus, Hemaglutinin) are not so obvious. Sequences in the 40%–20% sequence identity with acceptable alignment length (above 90%) are checked for similarity using the PDB website. In these cases and in general ones, Bit score is often a good indicator of where the cutoff should happen. The dramatic drops from one sequence to the next indicate a possible non-related structure.

The accepted sequences are then downloaded as a whole protein structure and split using `split.pdb()` function, producing single chain PDB files from multi-chain structures. Each chain sequence in the alignment is then aligned to reference sequence using the `pdbaln()` function. This function aligns sequences using MUSCLE and returns a matrix of aligned sequences from the downloaded PDB files.

Next, 10%, 5%, or 2% of the length of the reference sequence (or the longest sequence in the alignment) is calculated to assess the number of residues by which the sequences are allowed to differ (the fractions are rounded to the nearest whole number). The reference chain or the longest chain in the alignment is now added to the new alignment matrix. This chain is then compared to the next chain in the original alignment (gaps are avoided entirely in this part of the analysis). If the chains are different enough at aligned residues, a sequence is transferred from the

old alignment into the new one. If not, the next sequence in the old alignment is compared against the sequence(s) in the new alignment. Every time a chain is transferred from the old to the new alignment, the next sequence in the old alignment is checked against each sequence in the new alignment. Thus, the resulting new alignment contains chains that are different at given percentages of residues.

Once the final alignment is established relative solvent accessibility (RSA) is found. Firstly, solvent accessibility (SA) is determined using the DSSP function `dssp()`. (In the case of hemaglutinin, `dssp()` function does not work due to numbering of residues in this structure. A separate program ran to renumber those hemaglutinin structures that had this issue). This function is implemented to create a new matrix with all the values representing SA values for a given residue in its original location in the alignment matrix. SA values were then divided by theoretical normalizing values (CITING HERE) for each residue, creating a RSA matrix with RSA values located identically to its residue location in the alignment.

Next, each column in the RSA matrix is then averaged to produce a weighted mean RSA value at each location. The weights are calculated using phylogenetic trees produced by RAxML. The weighted mean RSA is now a vector with all previous column values averaged at all the non-gap positions.

Second part of the analysis involves finding structural differences among the aligned structures. Function `fit.xyz()` structurally aligns the PDB files producing new x, y, and z positions of the α-carbon atom in each aligned residue. The structures are positioned to get the least possible root mean square deviation

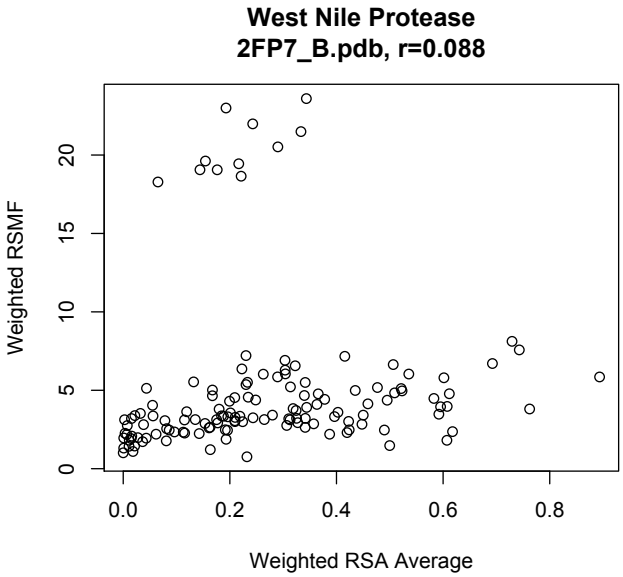
(RMSD) based equally on all non-gap positions throughout the structure. Then, the weighted root mean square fluctuation (RMSF) is calculated, using the same weights as previously mentioned. At this point a vector is created containing weighted variances for x, y, and z positions consecutively with respect to aligned residue positions. Distance formula is then applied to these values to make up one value.

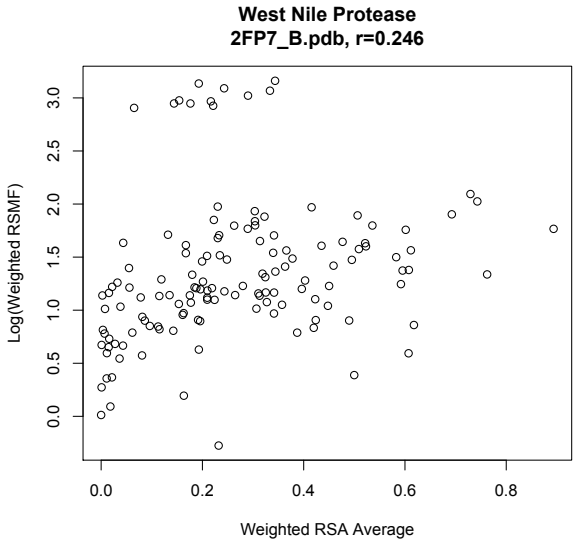
Viral Protein	PDB Structure	Chain	Total BLAST Hits	BLAST Hits Used	Unique Sequences with Difference[1]		
					10%	5%	2%
Dengue Protease-Helicase	2VBC	A	149	30	0 / 0%	7 / 23%	7 / 23%
		B	7	7	2 / 29%	3 / 43%	3 / 43%
West Nile Protease	2FP7	A	21	12	2 / 17%	2 / 7%	5 / 42%
		B	77	15	2 / 13%	6 / 40%	6 / 40%
HIV – 1 Reverse Transcriptase	2HMI	B	354	322	2 / 0.6%	2/ 0.6%	3 / 0.9%
Influenza Nucleoprotein	3TG6	A	18	5	2 / 40%	2 / 40%	3 / 60%
Marburg RNA binding domain	4GH9	A	44	32	3 / 9%	3 / 9%	3 / 9%
Hepatitis C Protease	4AEX	A	353	222	4 / 2%	5 / 2%	10 / 5%
Japanese Encephalitis Helicase/Nucleoside	2Z83	A	119	31	6 / 19%	7 / 23%	7 / 23%
Crimean-Congo Hemorrhagic Fever Nucleocapsid	4AKL	A	16	6	1 / 17%	1 / 17%	3 / 50%
Rift Valley Fever Virus Nucleoprotein	3OV9	A	145	73	1 / 1%	1 / 1%	3 / 4%
Hemaglutinin Precursor	1RD8	A	365	313	14 / 4%	19 / 6%	27 / 9%
		B	343	309	8 / 3%	9 / 3%	14 / 45%

[1]Results presented as count/percent of BLAST hits used.

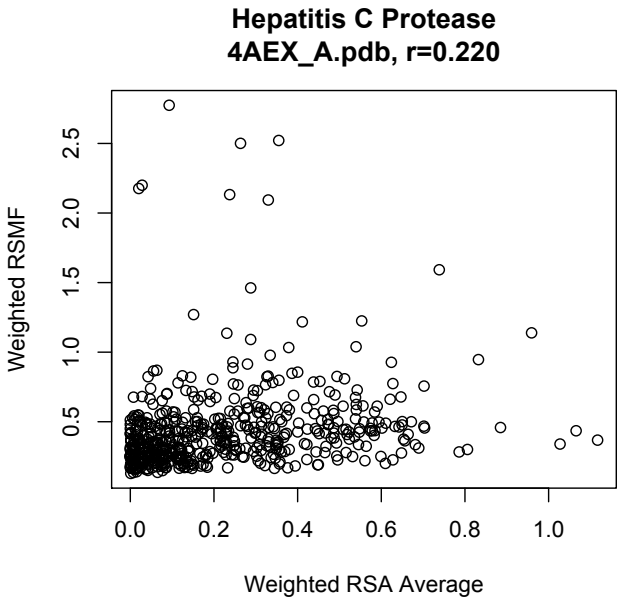
Table 1. Unique sequence found through BLAST in viral protein PDB structures. Numbers in BLAST hits used column are the sequences that pass ≥40% sequence identity and ≥90% alignment length criteria from the total found. Generally the number of unique sequences found is rather low compared to BLAST hits used; except in those cases with low number of BLAST hits used. The highest percentage of unique sequences found is mostly in those PDB structures with lowest number of used BLAST hits.

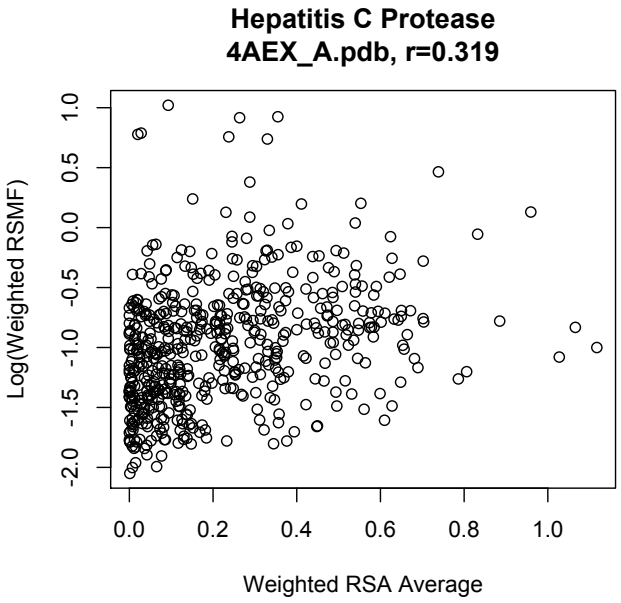
a) West Nile Protease Weighted RSA Average vs Weighted RMSF Plots.



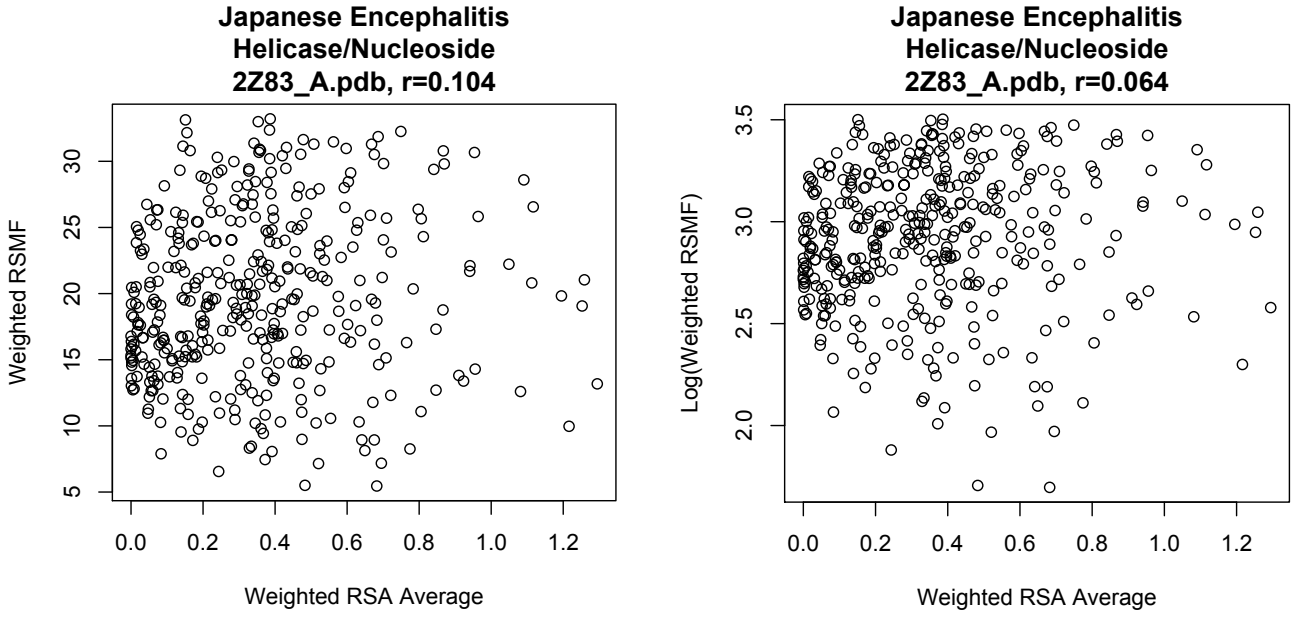


b) Hepatitis C Protease Weighted RSA Average vs Weighted RMSF Plots.





c) Japanese Encephalitis Helicase/Nucleotide Weighted RSA Average vs Weighted RMSF Plots.



d) Hemaglutinin Precursor Weighted RSA Average vs Weighted RMSF Plots.

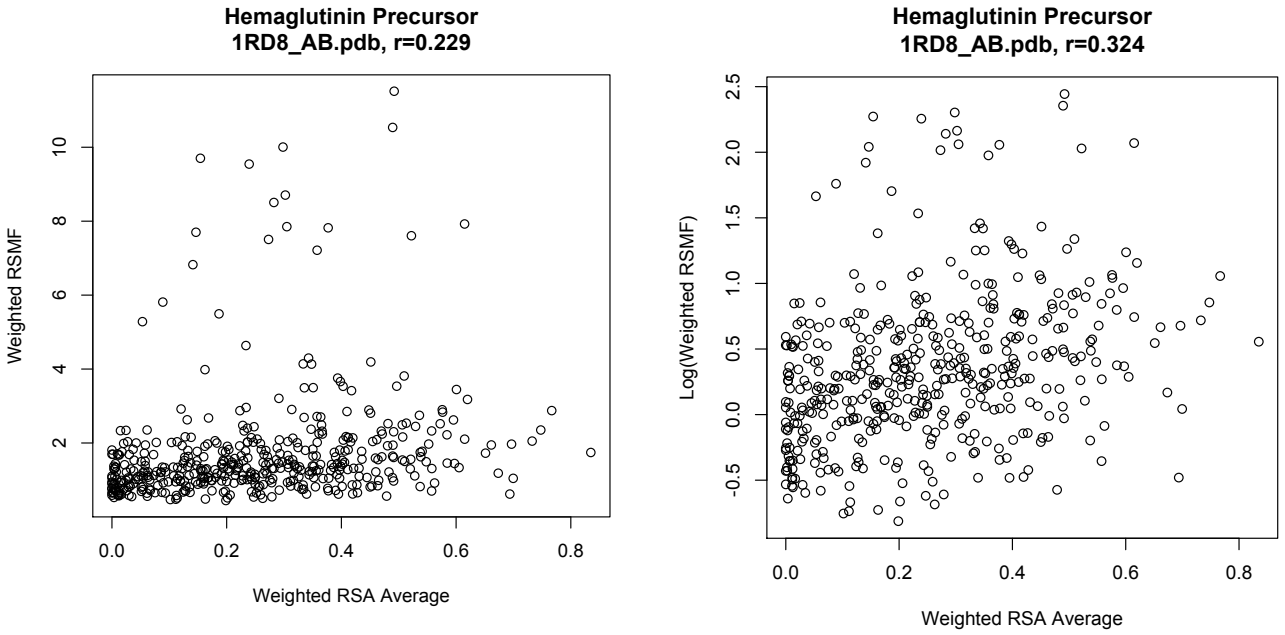


Figure 1. Average RSA and RMSF relationship in Dengue, West Nile, Hepatitis C, Japanese Encephalitis, and Hemaglutinin proteins. The relationships were adjusted

with log function. In all of the cases except for Japanese Encephalitis the log adjustment increased the correlation. For all others, the correlation between these two values is about 25-30%.