INTRODUCTION

- The Influenza Research Database (IRD, http://www.fludb.org) is a NIAID-funded, publicly available database resource that integrates biological data and bioinformatics analysis tools for the support of Influenza research.
- Here we describe a component of IRD called Sequence Feature Variant Type (SFVT) that describes regions called ‘Sequence Features’ (SF) within influenza proteins based on their structural (e.g. beta-sheet), functional (e.g. active sites), sequence alterations (mutations that cause altered phenotype) or immune epitope properties.
- SFs for influenza proteins have been defined from published literature (www.ncbi.nlm.nih.gov/pubmed) and from public domain databases such as UniProt (www.uniprot.org) and the Immune Epitope Database (IEDB, http://iedb.org)
- The size of a SF can range from a single amino acid residue to a complete protein; it can be present on any sub-region of the genomic sequence; it can be overlapping, continuous or non-contiguous.
- The extent of sequence variation is then be described as a collection of ‘Variant Types’ (VT) for each SF, computed by multiple sequence alignment of all relevant influenza sequences in IRD.
- SFVT correlation analysis can allow for the identification of genetic determinants of important phenotypic characteristics of the virus such as drug sensitivity/resistance, measures of virulence, host range restriction, etc.

OVERVIEW OF THE SFVT APPROACH

Table 1: Summary of the total SFs defined as of June 2011 in IRD for all influenza proteins. The column listing the total count of SFs for each flu protein is hyperlinked to list of all SFs for that protein.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Total SFs</th>
<th>Total Amino Acid Instanced</th>
<th>Amino Acid Pos.</th>
<th>Start Position</th>
<th>Total Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>200</td>
<td>500</td>
<td>200</td>
<td>1</td>
<td>500</td>
</tr>
<tr>
<td>N1</td>
<td>150</td>
<td>250</td>
<td>150</td>
<td>1</td>
<td>250</td>
</tr>
</tbody>
</table>

GUIDELINES USED IN NAMING SFVTs

The SF name is assigned using the following syntax, wherein each word is delimited by an underscore:

Influenza virus type_ protein symbol_ sequence feature type_ start position of the SF (total length of SF).

Example: Influenza_A_H1N1_cyttoplasmic_domain_550(16)

Sequence variations occurring for each SF are called “Variant Types” (VT) and can be computed by multiple sequence alignments of all relevant Influenza virus genomes in IRD.

The first VT (VT-1) corresponds to the specific sequence string found in the reference strain chosen for that protein

REPRESENTATION OF VARIANT TYPES IN IRD

Figure 1: A detailed report of the SF description for the NS1 nuclear export signal together with a table displaying the VTs identified by multiple sequence alignments (using the MUSCLE program) for the SF in the IRD database.

HIGHLIGHTING SFs ON PROTEIN 3D STRUCTURE

Figure 5: shows the NS1 protein effector domain (PDB ID: 3EE9); the residues highlighted in brown (amino acids 137-147) represents the Influenza_A_NSI_nuclear-export-signal_137(11) sequence feature.

STATISTICAL ANALYSIS OF SFVTs

- As an example use case, we have used the SFVTs of the NES region of NS1 to demonstrate the utility of this approach in sequence analysis and genotype-phenotype association studies that can in turn assist in making testable hypothesis for experimental validations.
- We analyzed the first 16 VTs of NES by looking at their distribution across broadly-classified host species groups: avian, human, equine, swine and others (which do not belong to the first five groups).
- We performed corrections for apparent geo-temporal bias using a weighting method that assumes the total number of viruses for all host groups to be uniform across all the years and for all regions of the world to have roughly the same flu prevalence. Thus, instead of counting each record as one, we used its new weights for further statistical tests.

Table 2: A partial list of SFs defined for NS1 protein shows SF ID, SF name, total number of VTs for each SF, amino acid start and stop positions, reference/publication and general comments about the SF

<table>
<thead>
<tr>
<th>Protein</th>
<th>Total SFs</th>
<th>Total Amino Acid Instanced</th>
<th>Amino Acid Pos.</th>
<th>Start Position</th>
<th>Total Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>200</td>
<td>500</td>
<td>200</td>
<td>1</td>
<td>500</td>
</tr>
<tr>
<td>N1</td>
<td>150</td>
<td>250</td>
<td>150</td>
<td>1</td>
<td>250</td>
</tr>
</tbody>
</table>

Figure 6: Bar graphs showing proportions of individual virus isolates from 6 different host species represented across 16 VTs of the Influenza_A_NSI_nuclear-export-signal_137(11) sequence feature type. (A) shows the distribution of VTs for unprocessed data, while figure (B) represents the distribution of hosts for the processed data.

We performed chi square tests on the SFVTs across the 6 different host categories to calculate the p-value (see below), all of which turned out to be extremely small indicating that there is high skewing in the distribution of VTs across the different hosts despite controlling for geo-temporal bias, thereby indicating its possible role in host range phenotype.

Further laboratory experiments are underway to validate the observed significance of the NES SFVTs on host range restriction phenotype of the virus. Similar tests are also planned for analyzing other SFs of NS1 to determine their role in influencing this phenotype of the virus.

Table showing the changes in p-values for different VTs before (table A) and after (table B) controlling for geo-temporal bias in data.