INTRODUCTION

The Influenza Research Database (IRD) (http://www.fluadb.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 the Sequence Feature Variant Type (SFVT) component of IRD that delineates regions, called ‘Sequence Features’ (SF) within influenza proteins based on their structural (e.g. beta-sheets), functional (e.g. sialic acid binding) or immune epitope properties.

There is no restriction on the size of the SF and they can be present on virtually any sub-region of the genomic sequence. SFs can be overlapping, continuous or non-contiguous.

The extent of sequence variation is described as a collection of ‘Variant Types’ (VT) for each SF, computed by multiple sequence alignment of all relevant Influenza sequences in IRD.

Defining SFVTs will aid in identifying genetic determinants of important phenotypic characteristics, such as drug sensitivity/resistance, measures of virulence, host range restriction, etc.

Here we demonstrate how SFVT data can be analyzed and visualized using statistical methods and graphical representations.

OVERVIEW OF THE SFVT APPROACH

Table 1: Summary of the total number of SFs listed in the current beta version available on IRD for all Influenza proteins. The column listing the total number of sequence features for each Influenza segment is hyperlinked to list of those SFs for those proteins.

GUIDELINES USED IN NAMING SFVTs

The SF name is assigned using the following approach:

Influenza virus type _protein symbol_ SF type_ start position of the SF (total no. of amino acids that make up the SF)

Example: Influenza_A/H1N1_ cytoplasmic-domain_550(16)

For a given SF, each variant type (VT) is numbered in series and the number is indicated just after the length as shown:

Example: Influenza_A/H1N1_ cytoplasmic-domain_550 (16)_VT1

Table 3: A detailed report of the SF description for the NS1 nuclear export signal together with a table displaying all the variant types computed for the SF

STATISTICAL ANALYSIS OF SFVTs

As a preliminary analysis we have started performing genotype-phenotype associated statistical analysis of the SFVTs of the Nuclear Export Protein (NEMP) of Influenza non-structural protein 1(NS1) to look for potential NS1-associated host-range restriction. The data shows high skewing in distribution of VTs across country of isolation, host, year and serotype.

Fig: The bar graph, shows the proportion of individual virus isolates from a given host species group that carry particular VTs for the InfluenzaA_NSI_nuclear-export-signal_137(10).

Several interesting observations can be made regarding the potential role of the nuclear export signal in host-range restriction.

- We performed the Chi square and Fisher exact tests on the SFVTs across six different host categories (Avian, Chicken, Human, Equine Swine and others) to calculate the P-value, all of which have turned out to be extremely small indicating that there is high skewing in the distribution of VTs across the different hosts.

- The tests also show which host is responsible for the skewing. Example: Some VTs are restricted only to human hosts (VT-9) while some are found predominantly in horse (VT-10) over a wide spatial-temporal range and some VTs are mostly restricted to avian and/or chicken (VT-14, VT-16) host groups.

The table below shows the small P-values obtained for the SFVTs of the nuclear export protein

FUTURE PLANS:

Our observations thus far pose interesting questions about the potential causes for the restriction of certain VTs to specific host groups.

We are working on performing more rigorous tests to show co-relation between the NS1 protein and host-range restriction of Influenza that could potentially be proved by laboratory experiments.