BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 9(8)

Design of a set of probes with high potential for influenza virus epidemiological surveillance

Luis R Carreño-Durán[1], V Larios-Serrato[1], Hueman Jaimes-Díaz[1], Hilda Pérez-Cervantes[1], Héctor Zepeda-López[2], Carlos Javier Sánchez-Vallejo[1], Gabriela Edith Olguín-Ruiz[1], Rogelio Maldonado-Rodríguez[1] & Alfonso Méndez-Tenorio[1]*

[1]Laboratory of Biotechnology and Genomic Bioinformatics, Department of Biochemistry, Escuela Nacional de Ciencias Biológicas, Instituto Politécnico Nacional, México City, Méxic; [2]Laboratory of Conservation Medicine, Escuela Superior de Medicina, Instituto Politécnico Nacional, México City, México; Alfonso Méndez-Tenorio – Email: amendezt@ipn.mx; Phone: (55) 57296000 ext. 62322; *Corresponding author

Received April 09, 2013; Accepted April 10, 2013; Published April 30, 2013

Abstract:
An Influenza Probe Set (IPS) consisting in 1,249 9-mer probes for genomic fingerprinting of closely and distantly related Influenza Virus strains was designed and tested in silico. The IPS was derived from alignments of Influenza genomes. The RNA segments of 5,133 influenza strains having diverse degree of relatedness were concatenated and aligned. After alignment, 9-mer sites having high Shannon entropy were searched. Additional criteria such as: G+C content between 35 to 65%, absence of dimer or trimer consecutive repeats, a minimum of 2 differences between 9mers and selecting only sequences with Tm values between 34.5 and 36.5oC were applied for selecting probes with high sequential entropy. Virtual Hybridization was used to predict Genomic Fingerprints to assess the capability of the IPS to discriminate between influenza and related strains. Distance scores between pairs of Influenza Genomic Fingerprints were calculated, and used for estimating Taxonomic Trees. Visual examination of both Genomic Fingerprints and Taxonomic Trees suggest that the IPS is able to discriminate between distant and closely related Influenza strains. It is proposed that the IPS can be used to investigate, by virtual or experimental hybridization, any new, and potentially virulent, strain.

Keywords: IPS, fingerprinting, Virtual Hybridization, Shannon Entropy, Microarray, Influenza virus.

Background:
Influenza viruses are part of *Orthomixoviridae* Family and possess segmented genomes consisting of seven or eight separate RNA molecules, each coding for one or more viral proteins. The viruses can exchange segments, leading to diversity of reassortant strains. Together with accumulation of point mutations, segment reassortment is the basis for evolution and maintenance of diversity for these viruses. It provides them with the ability to rapidly adapt to the pressure of the host immune system and leads to the continuous emergence of new virus variants that cause seasonal and pandemic outbreaks of influenza. Because of this ability, segmented viruses can exist in numerous genotypes and serotypes, presenting a challenge to the creation of protective vaccines and detection methods **[1, 2].**

Because of these reasons, the early detection and diagnostic confirmation of influenza virus infections is fundamental for an appropriate control of the disease. Several molecular biology techniques, most of them based on PCR amplification, have contributed to the diagnostic of the different types and subtypes of influenza virus. However, PCR techniques are frequently unable to detect new potentially virulent strains. Other techniques such as sequencing are able to perform a precise

identification of such strains but still are not so widely available for routinary diagnostic **[3, 4].**

The creation of a microarray is complicated when genomic structures are similar. Probe selection is further complicated when the number of known sequences is very large. When this happens the probe selection strategy becomes critical **[5].** There are several methods **[6-12]** for the selection of specific probes for influenza virus detection. Direct search for probes based on traditional computational methods is labor-intensive and often requires plenty of time. The Shannon entropy (*H*), is a bioinformatics technique that has been used to sort the influenza virus, to analyze the evolution of influenza **[13]**, to facilitate the development of an anti-influenza vaccine **[14]**, and to create a profile of these areas of high variation, observing characteristic patterns for each subtype **[15]**.

In the present approach we designed and tested *in silico*, an Influenza Probe Set (IPS) which consists in 1,249 probes with a length of 9mer, extracted from sequence alignment zones with maximum entropy within the full viral genome of over 5,000 viruses reported, considering almost all viral subtypes of Influenza A. Using Virtual Hybridization (VH) technology, *in silico* Genomic Fingerprints were generated, which in turn were compared to estimate a phylogeny based on the fingerprint pairwise distances. Other studies have employed the use of the VH technology to create genomic fingerprints for *in silico* classifying of microorganisms as Human Papillomaviruses **[16]** and bacterial genomes **[17].**

Methodology:

Shannon entropy is a measure of the lack of predictability of an element **[19]**, such as a given base, in a particular position of alignment. Highly variable columns in an alignment will yield maximum values of entropy.

Search Probe

This program developed in Java, calculates the Shannon entropy of aligned sequences. It finds the points having maximum entropy, then, selects 9-mer sequences (the size can be modified by the user), using the point of maximum entropy as the 9-mer center.

The equation used by SearchProbe to calculate the Shannon entropy is: $H(n) = -\Sigma f(i, n) \ln f(i, n)$; Where H (n) is the entropy at position *n*, *i* represents a residue (in this case there are only four possible options A, C, G and U), $f(i, n)$ is the frequency of residue *i* in the n position. The information content in position *n*, is defined then as a decrease in uncertainty or entropy in that position. In our particular case, SearchProbe seeks regions with maximum entropy values **[18].**

CalcProbes. This Perl script refines the search of probes using the 9-mer sequences provided by SearchProbe. These sequences are subject to the next restrictions: i) Select only sequences having between 35-65 %G + C (4 or 5), ii) Eliminate 9-mers having tandem repeats of 2 or 3 nucleotides, iii) Select sequences having a minimum of 2 differences between them and iv) Chose 9-mer sequences having 34 to 36°C Tm values. Tm values were calculated with the thermodynamic Nearest-Neighbors (NN) model using SantaLucia parameters **[19].** The final 1,249 9-mer probe set selected by this procedure is the IPS (Influenza Probe Set).

Virtual Hybridization (VH)

Virtual Hybridization is a computer program able to predict perfect and mismatched target/probe hybridizations under a selected Tm cutoff value. The stability of target/probe duplexes is calculated with the NN model. This program was used to determine all the hybridizations occurring between each Influenza virus genome, or control strain, and the IPS. The group of hybridization signals produced by each viral genome corresponds to its particular fingerprint **[20].**

Genomic Fingerprinting Analysis with UFA software

Universal Fingerprinting Analysis (UFA) software transforms genomic fingerprints produced by Virtual Hybridization under any chosen stringent condition, into images. It also allows visual comparison of any selected pairs of fingerprints, producing spots with specific colors for both distinctive as well as for shared hybridization signals. Besides, this tool is able to calculate pairwise distances between pairs of genomic fingerprints. From a table of such distances Taxonomic trees were built using the Neighbor -Joining method with the program MEGA 5 **[21].**

Distinction of Influenza strains with the IPS

Two types of analysis were performed: I) A Taxonomic tree, based on distances between IPS-Genomic Virtual Hybridization fingerprints, comparing several types of Influenza and other viruses, was made. II) Overlapped images from selected pairs of genomic fingerprints for strains having: low, medium, or high degree of relatedness, were made. *Influenza A /mallard duck/New York/170/1982(H1N2)* and *Influenza A/Mexico/InDRE4487/200* were used as references.

Results & Discussion:

In the first step an average of 550,500 non-unique sequence probes were selected from the alignment. Furthermore probe sequences were clustered in order to remove the repeated ones and to select only those with entropy higher than a convenient threshold (ProbeSearch). Calcprobes is responsible for applying the design parameters explained in the methodology. After the above-mentioned, we performed a third selection, by removing sequences containing probes with the lowest entropy values and taking probes with a Tm range of 34.5 to 36.5°C and free energy values between -9.00 and -13.5Kcal/mol.

Virtual Hybridization

A database of tested target viral genomes used for the *in silico* experiments was created. The VH programs conducts a rigorous and reliable analysis to find and track all the sites in each viral genome where the probe sequences can hybridize taking into account the degree of complementarity between the probe and the recognized site in the target (allowing at least a mismatch difference) and the thermodynamic stability between them. The generated information constitutes an *in silico* genomic fingerprint listing details of the specific sites in each target DNA where hybridization occurred, the number and sequence of the probe that hybridized as well as the free energy value of the hybridizations and it also provides the sequence of the target site recognized by each probe. A free energy cutoff value of -9 kcal/mol for 9mer probes was used.

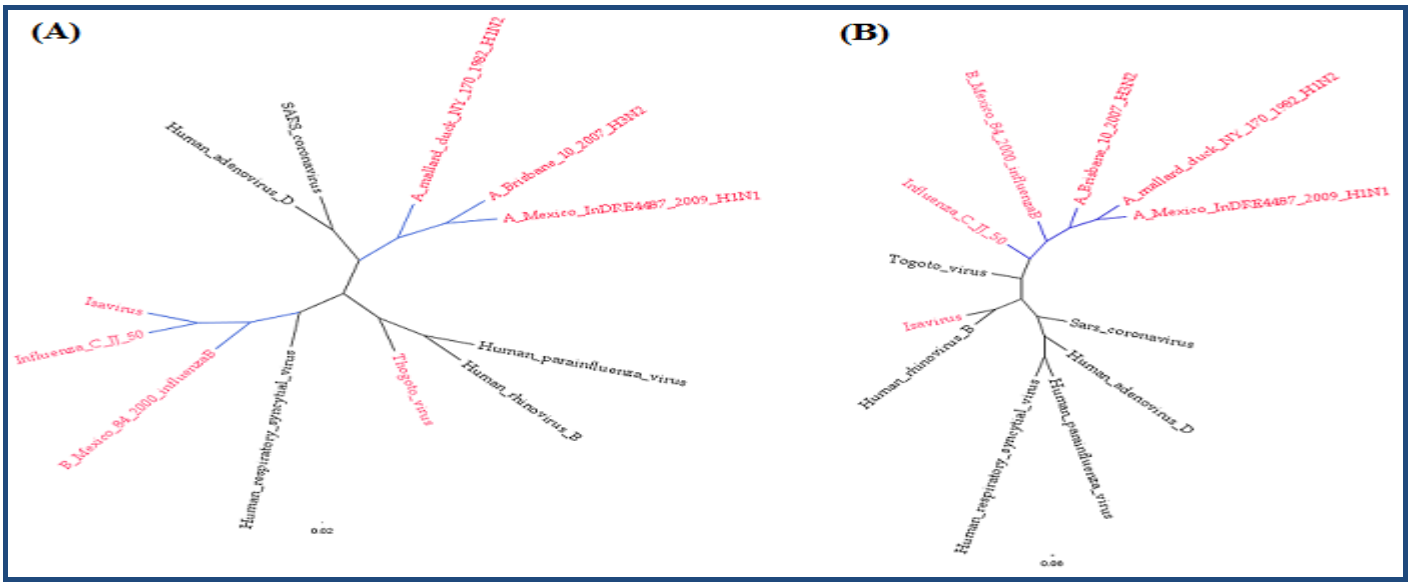

Figure 1: Taxonomic trees of 12 viral families including *Paramixoviridae, Orthomixoviridae, Coronaviridae, Picornaviridae, Adenoviridae,* Influenza A (H1N1, H1N2, H3N2), B and C, and two other Orthomixovirus, Thogotovirus and Isavirus is given (in red). **(A)** Fingerprinting Tree, **(B)** Alignment Tree. It is shown that all the Influenza A virus subtypes were clustered into a single group.

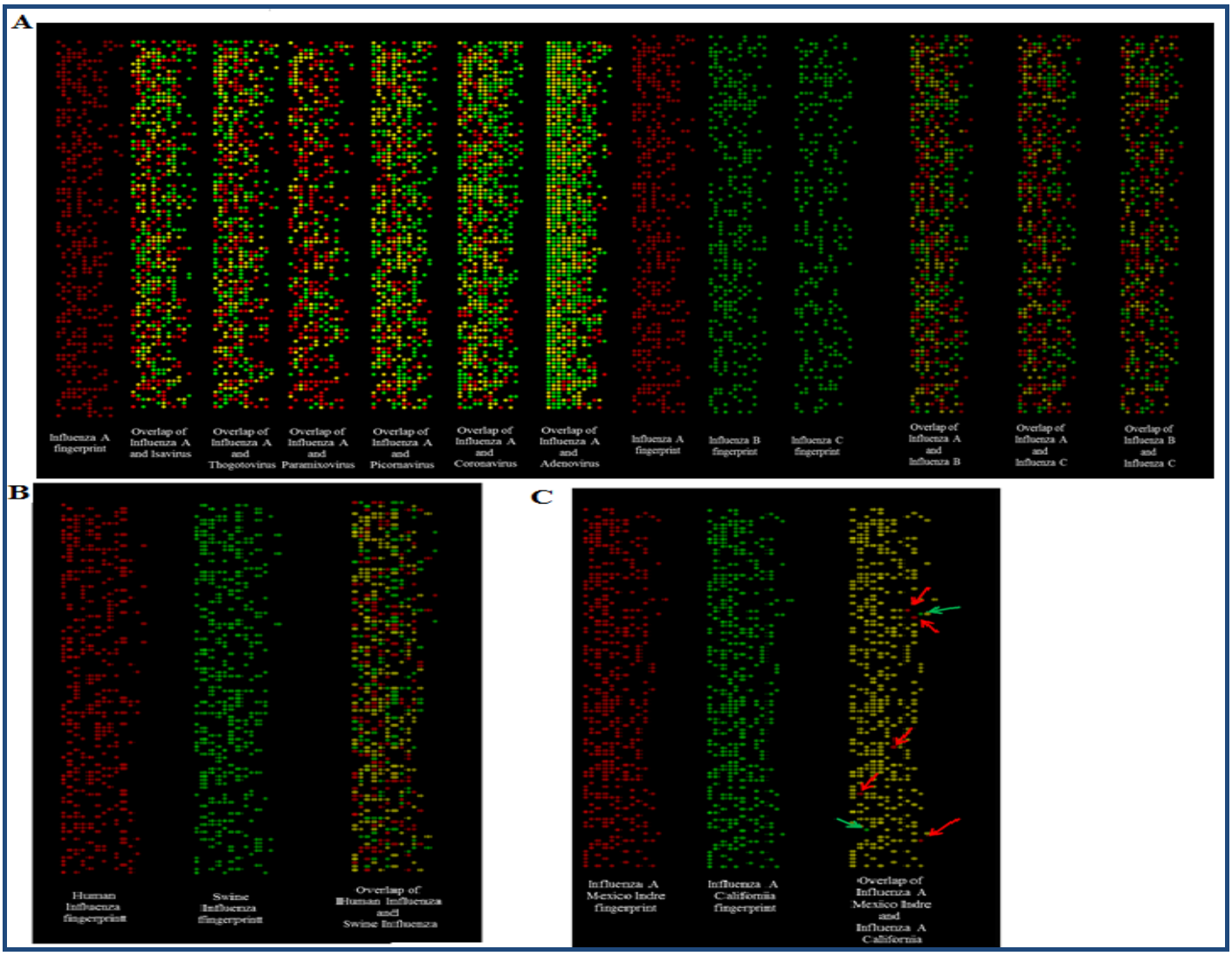

Figure 2: A) Genomic fingerprints of different influenza viruses and other viral families. Using as reference organism the virus Influenza A *A /mallard duck/New York/170/1982(H1N2)* (in red) and the Infectious salmon anemia virus(Isavirus), Thogotovirus,

Human respiratory syncytial virus*(Paramixoviridae)*, Human rhinovirus B (*Picornaviridae*), SARS coronavirus (*Coronaviridae*), Human adenovirus D (*Adenoviridae*) in green to compare the fingerprints generated **,** Genomic fingerprints of different viral types of influenza virus. Using as reference organism the virus Influenza A *A /mallard duck/New York/170/1982(H1N2)* (in red) and Influenza B *B/Mexico/84/2000* and Influenza C *C/Ann Arbor/1/50* (in green) to compare fingerprints **B)** Genomic fingerprints of different viral types of influenza virus. Using as reference organism the virus *A/New York/18/2006/H1N1* (in red) and *A/ Swine/Wisconsin/1915/1988/H1N1* (in green) to compare fingerprints; **C)** Genomic fingerprints of different viral types of influenza virus. Using as reference organism the virus *A/Mexico/InDRE4487/2009(H1N1* (in red) and *A/California/04/2009 H1N1* (in green) to compare fingerprints.

Evaluation of the genomic fingerprint

The analysis of the performance of our set of probes to distinguish between several viral sequences the *in silico* evaluation was divided into several steps: **A)** *Viral Family Tes*t: Two kinds of trees were calculated: One derived from the alignment of the concatenated fragments calculated with Clustal X 2.0 and the other resulting from the comparison of genomic fingerprints obtained with the IPS and VH. Both trees are derived from 12 different viral genome families including *Paramyxoviridae, Coronaviridae, Picornaviridae, Adenoviridae*, which cause respiratory symptoms similar to influenza and several family *Orthomixoviridae* viruses like influenza B, influenza C, influenza A (H1N1, H1N2, H3N2) Thogotovirus and Isavirus are shown in **(Figure 1)**.

The comparison of the two trees shows a close correspondence. The tree from genomic fingerprints groups the influenza A (H1N1, H1N2 and H3N2) on a branch, influenza B, C and Isavirus in another branch and other viral families in other clusters. It is noteworthy that Human Respiratory Syncytial virus (which is a Paramyxovirus) was grouped with the Rhinovirus Human Rhinovirus B, together with Thogotovirus in the tree based on genomic fingerprints. Thogotovirus has been classified as belonging to the Orthomixovirus family, although other studies make comparisons of *Orthomyxoviridae* PB1 proteins, showing low percentages of amino acid identity when compared with influenza viruses and Isavirus **[22, 23].** Likewise, influenza virus types A, B and C yielded characteristic patterns for each virus, so IPS probes allowed creating distinctive fingerprints for each one and create one fingerprint characteristic for each virus **(Figure 2)**.

B) Hybridization of viral genomes of the same subtype was carried out on two subtypes of Influenza A H1N1, that infect different hosts (human and swine), to check if the IPS is able to generate distinctive genomic fingerprints. The virus *A/New York/18/2006/H1N1* causes seasonal influenza whereas the virus *A/ Swine/Wisconsin/1915/1988/H1N1* infects swine. It is highlighted that the IPS probes generated specific genomic fingerprints for each one. This result is very relevant showing that IPS is capable of an appropriated identification when there are outbreaks of this disease in humans by strains from animals such as pigs **(Figure 2)**.

C) Comparison of Genomic Fingerprints of two genomes with very high similarity. Overlapping Genomic Fingerprints of Influenza A H1N1 viruses *A/Méxicoindre4487/2009* and A*/California/04/2009* from the 2009 pandemics are shown in Figure 2. It is clear that both viruses are very similar with only minor mutations, as expected for viruses from the same outbreak. However IPS genomic fingerprints are able to show seven differences between then, with five specific probes for *A/Méxicoindre4487/2009 H1N1* virus and two for the *A/California/04/2009*. This is very important for molecular studies of influenza because IPS is highly sensitive as to spread viruses even those very closer; this will help in the management of influenza epidemiology, and not depend on a previous sequencing.

Conclusions:

Following the established parameters, the set of 1249 highly specific probes (IPS) allowed us to correct typing and subtyping of influenza viruses, including human and animal strains, as well as very similar strains. The IPS design based on the construction of probes from regions of the viral genome with maximum entropy allows a highly sensitive discrimination.

Through an *in silico* hybridization, the performance of the IPS microarray was simulated, allows us to know the possible behavior of the probes, and predicting genomic fingerprints of these viruses. Prediction is based in experimentally supported thermodynamic models, which suggest that the IPS microarray would be a valuable Influenza diagnosis tool.

Acknowledgement:

Authors would like to thank to Engineer Cesar Arturo Zapata Acevedo for their contributions to this work

References:

[1] McHardy AC & Adams B, *PLoS Pathog*. 2009 **5**: 10 [PMID: 19855818]
[2] Steinhauer DA *et al. Annu Rev Genet*. 2002 **46**: 97 [PMID: 22934646]
[3] Writing Committee of the WHO Consultation on Clinical Aspects of Pandemic (H1N1) 2009 Influenza, *N Engl J Med.* 2010 **362**: 1708 Erratum in: *N Engl J Med*. 2010 **362**:2039 [PMID: 20445182]
[4] Ryabinin VA *et al*. *PLoS One*. 2011 **6:** 4 [PMID: 21559081]
[5] Sengupta S *et al. J Clin Microbiol*. 2003 **41**: 10 [PMID: 14532180]
[6] Li J *et al. J Clin Microbiol*. 2001 **39**: 2 [PMID: 11158130]
[7] Relogio A *et al. Nucleic Acids Res*. 2002 **30**: 11 [PMID: 12034852]
[8] Fesenko EE *et al*. *Influenza Other Respi Viruses*. 2007 **1**: 3 [PMID: 19453417]
[9] Dawson ED *et al*. *Anal Chem*. 2006 **78:** 22 [PMID: 17105150]
[10] Townsend MB *et al. J Clin Microbiol*. 2006 **44**: 8 [PMID: 16891504]
[11] Kessler N *et al. J Clin Microbiol*. 2004 **42**: 5 [PMID: 15131186]
[12] Li H *et al. J Clin Microbiol*. 2007 **45**: 7 [PMID: 17507510]
[13] Pavesi A, *Gene*. 2007 **402**: 1 [PMID: 17825505]
[14] Heiny AT *et al*. *PLoS One*. 2007 **2**: 11 [PMID: 18030326]
[15] Thompson WA & Fan S, *Entropy.* 2008 **10**: 736doi:10.3390/e10040736
[16] Alfonso M *et al. Rev Latinoam Microbiol*. 2006 **48**: 2 [PMID: 17578073]

[17] Hueman JD *et al. J Microbiol Methods*. 2011 **87**: 3 [PMID: 21906634]
[18] Shannon CE, *Techn J*. 1948 **27**
[19] SantaLucia J, *Jr Proc Natl Acad Sci USA*. 1998 **95**: 4 [PMID: 9465037]
[20] Casique-Almazán *et al. Bioinformation*. 2012 **8**: 12 [PMID: 22829736]
[21] Tamura K & Kumar S, *Mol Biol Evol*. 2011 **28** :10 [PMID: 21546353]
[22] Bjørn Krossøy & Curt Endresen, *J Virol*. 1999 **73**: 3 [PMID: 9971796]
[23] Cottet L & Cortez-San Martin M, *J Virol*. 2010 **84**: 22 [PMID: 20810724]

Edited by P Kangueane
Citation: **Durán *et al.*** Bioinformation 9(8): 414-420 (2013)

BIOINFORMATION

open access

Supplementary material:

Table 1: Table of distances pairwaise score generated in VH

	A/mallard /duck/NY/ 170/1982/ H1N2	*A/Brisban e/10/2007/ H3N2*	*Human/ adenovir us/D*	*Human/ parainfl uenza/vi rus*	*Human/ rhinovir us/B*	*B/Mexic o/84/200 0/influen zaB*	*Influenz a/C/JJ/50*	*Isavir us*	*Human/ respirator y/syncytia l/virus*	*A/Mexico/ InDRE448 7/2009/H1 N1*	*Tho goto viru s*	*SARS/ Coron avirus*
A/mallard/du ck/NY/ 170/1982/H1 N2	0											
A/Brisbane/1 0/2007/ H3N2	0.097617	0										
Human/adeno virus/D	0.123168	0.126535	0									
Human/parai nfluenza/viru s	0.162238	0.160721	0.131612	0								
Human/rhino virus/B	0.175138	0.1846	0.183593	0.179549	0							
B/Mexico/84/ 2000/ influenzaB	0.150161	0.148574	0.13281	0.154635	0.173942	0						
Influenza/C/JJ /50	0.165953	0.164234	0.157487	0.176905	0.198252	0.156114	0					
Isavirus	0.152127	0.14312	0.131786	0.160429	0.199084	0.145391	0.157201	0				
Human/respir atory/ syncytial/viru s	0.128253	0.139808	0.106539	0.129708	0.177608	0.125081	0.139196	0.133954	0			
A/Mexico/In DRE4487/20 09/H1N1	0.091123	0.072992	0.120571	0.156324	0.187606	0.136418	0.161303	0.138512	0.129872	0		
Thogotovirus	0.153444	0.158864	0.129435	0.151038	0.182098	0.155691	0.167459	0.160619	0.134631	0.158375	0	
SARS/Corona virus	0.141617	0.141085	0.09959	0.141824	0.182795	0.136951	0.150484	0.135656	0.118256	0.137504	0.151378	0

Table 2: Viruses accession number, names and viral family

Accession Number	Virus	Viral Family
Segment1 CY014908 Segment2 CY014907 Segment3 CY014906 Segment4 CY014901 Segment5 CY014904 Segment6 CY014903 Segment7 CY014902 Segment8 CY014905	A/mallard duck/New York/170/1982(H1N2)	Influenza A(*Orthomixoviridae*)
Segment1 CY035029 Segment2 CY035028 Segment3 CY035027 Segment4 CY035022 Segment5 CY035025 Segment6 CY035024 Segment7 CY035023 Segment8 CY035026	A/Brisbane/10/2007(H3N2)	Influenza A(*Orthomixoviridae*)
Segment1 FJ998206 Segment2 FJ998226 Segment3 FJ998223 Segment4 FJ998208 Segment5 FJ998217 Segment6 FJ998214 Segment7 FJ998211 Segment8 FJ998220	A/Mexico/InDRE4487/2009(H1N1)	Influenza A(*Orthomixoviridae*)
Segment1 FJ969516 Segment2 GQ377049 Segment3 FJ969515	A/California/04/2009 H1N1	Influenza A(Orthomixoviridae)

Segment4 GQ117044 Segment5 FJ969512 Segment6 FJ969517 Segment7 FJ969513 Segment8 CY053273 Segment1 CY018451 Segment2 CY018452 Segment3 CY018450 Segment4 CY018445 Segment5CY018448 Segment6 CY018447 Segment7 CY018446 Segment8 CY018449	B/Mexico/84/2000	Influenza B(*Orthomixoviridae*)
Segment1 AB126191 Segment2 AB126192 Segment3 AB126193 Segment4 AB126194 Segment5 AB126195 Segment6 AB126196 Segment7 AB283001	C/Ann Arbor/1/50	Influenza C(*Orthomixoviridae*)
Segment1 CY039924 Segment2 CY039923 Segment3 CY039922 Segment4 CY039917 Segment5 CY03992 Segment6 CY039919 Segment7 CY039918 Segment8 CY039921	A/ Swine/Wisconsin/1915/1988/(H1N1)	Influenza A(*Orthomixoviridae*)
Segment1 CY025236 Segment2 CY025235 Segment3 CY025234 Segment4 CY025229 Segment5 CY025232 Segment6 CY025231 Segment7 CY025230 Segment8 CY025233	A/New York/18/2006/(H1N1)	Influenza A(*Orthomixoviridae*)
Segment1 56407674 Segment2 56407672 Segment3 56407670 Segment4 56418533 Segment5 56407668 Segment6 56407666 Segment7 56403993 Segment8 56403990	Infectious salmon anemia virus	Isavirus (*Orthomixoviridae*)
Segment1 56403988 Segment2 56403977 Segment3 56403979 Segment4 56403984 Segment5 56403986 Segment6 56403981	Thogoto virus	Thogotovirus (*Orthomixoviridae*)
9629198	Human respiratory syncytial virus	Pneumovirus (*Paramixoviridae*)
9626735	Human rhinovirus B	Rhinovirus (*Picornaviridae*)
30271926	SARS coronavirus	Coronavirus (*Coronaviridae*)
190340974	Human adenovirus D	Adenovirus (*Adenoviridae*)