International Journal of Scientific Research & Engineering Technology (IJSET)
Vol.7 pp. 1-7

New *In Silico* Approach for the Determination of the Genetic Factors Associated with the Virulence of HxNy Influenza-A Family

Rima Soli[#1], Safa Berraies[#], Belhassen Kaabi[#2], Chokri Maktoof[*], Mourad Barhoumi[**],

and Sami Ben-Hadj Ahmed[***]

[#]*Pasteur Institute of Tunis, Tunisia.*
13 Place Pasteur, BP 74, 1002 Belvedere-Tunis, Tunisia.
[1]souli.rima@gmail.com

[3]belhassen.kaabi@gmail.com

**Laboratory of Biophysics and Nuclear Medicine*
Pasteur Institute of Tunis, Tunisia

***Laboratory of Molecular Epidemiology and Experimental Pathology,*
Pasteur Institute of Tunis, Tunisia.

****National Institute of Applied Sciences and Technology (INSAT), Tunis, Tunisia.*

***Abstract*— Capabilities of infiltration, replication and transmission of influenza-A virus to and from humans pose an imminent epidemic to pandemic threat to the world population and a major human healthcare burden. The viral genomic determinants, which facilitate these processes, are not well understood. Using Meta-data on influenza-A virus sequences and their respective estimated of basic reproductive number (R0) of different sub-types, we identified several genomic regions i.e. conserved patterns (using MEME) and SSR (using MiSA), by regressing R0 values against motif numbers and SSR repeats. Thus, we were able to detect genomic regions associated with virulence of the virus. A docking study showed also the association of the docking energy of HA and the sialic acid receptor and the number of repeats of some of these predetermined motifs (regions). Ninety-six (96) sequences of Influenza-A virus and their estimated R0 values covering H1N1, H3N2, H7N7, and H7N1 subtypes were acquired from the NCBI database. Using statistical linear regression, we identified several genomic regions in segments encoding the internal proteins (PB2, PB1, PA, M, NS) that are implicated in the virulence of VIAs, and these are correlated with the literature where they are already described.**

***Keywords*— Influenza-A, R0, MEME, MISA, virulence**

I. INTRODUCTION

Influenza A virus (IAV) is an influential pathogen causative of frequent epidemics and occasional pandemics in human. IAV pandemics remain the greatest infectious disease outbreaks in the past century.

Three main factors sway the diffusion potential of an influenza virus: (i) the ability to cause human disease, (ii) the herd immunity of the population to the virus, (iii) the replication and the transmission potential of the virus (virulence) [1]. At the virus level, the abilities to enter, and to replicate within the host cell, amplifying viral numbers, and thus, the potential for host-to-host transmission are the main determinant of virulence. This process requires multiple rounds of entering the cells, replications, virion assembly, and release. The assembly of IAV involves packaging of several host and viral proteins from an eight distinct segmented genome. Even though, the selective assembly of the eight-segment core remains one of the most interestingly unresolved problems in virology, the interaction between the IAV viral ribonucleo-protein (vRNP) complex and other host factors are major determinants of viral pathogenicity [2]. Type A viruses that affect mammals and birds are further classified into subtypes based on their 18 hemagglutinin (HA) and 11 neuraminidase (NA), which gives theoretically 198 (18x11) possible subtypes [3]. Not all subtypes of IAV infect humans and cause disease, however, many of them do.

Pandemics have been caused by subtypes H1N1 (1918, 2009), H2N2 (1957) and H3N2 (1968), and currently H1N1 and H3N2 are the circulating seasonal influenza A subtypes [4], [5]. H5N1 avian influenza viruses have caused the deaths of nearly 60% of humans that they have infected since 1997 and clearly represent a threat to public health [6]. H9N2 virus circulates widely in poultry, and has been responsible for sporadic human infections, in several regions. Few studies have been conducted on the pathogenicity of H9N2 isolates, that have different genomic features [7].

In addition, many laboratories have-confirmed H7N9 virus human infection cases have been recorded, with a case fatality rate of more than 30%. Clinical research has shown that cytokine and chemokine dysregulation contributes to the pathogenicity of the H7N9 virus [8]. Poultry exposure is a major risk factor for human H7N9 zoonotic infections, for which the mode of transmission is unclear [1].

ISSN 1737-9296

HA and NA proteins are used in the nomenclature of the virus subtypes. The HA protein complex mediates virus entry by binding to cell surface receptors and fusing the viral and endosomal membranes following uptake by endocytosis.

This poly- cleavage site of the HA proteins is considered one of the most important determinant contributing to the virulence of the IAVes.

On the other hand, viral NA is a surface protein of influenza virus that enables the virus to be released from the host cell surface. NAs are enzymes that cleave sialic acid groups from glycoprotein's and are also required for the IAV replication. However, evidence is now accumulating that these sites alone are not sufficient to establish the high virulence (pathogenicity), and that other sites located outside the HA protein cleavage spot, which are expressed by IAV that contribute to its pathogenicity.

As virulence is also in most part is determined by the host response, understanding the key host molecular driver(s) of virus-mediated disease, in relation to the viral genes, is also, a promising approach to host-oriented drug efforts in preventing disease.

On the population level, when an outbreak occurs, such as epidemic or pandemic influenza, it is necessary to provide criteria to characterize the dynamics of the disease within the population as well as its severity and virulence [9]. The basic reproduction number R0 is the main epidemiological parameter characterizing disease severity and virulence.

The basic reproductive number, R0is defined as the expected number of secondary cases produced by a single infection in a completely vulnerable population [10].

Based on the assumption that any genetic region influencing the virulence will be associated (correlated) with the estimated R0, the aim of this work is to identify, motifs (genetic regions) that are responsible for the virulence of the IAV' subtypes. This can be done, by regressing the number of repeats of these conserved regions (determined using software of pattern and motif detection), while assuming certain genetic homogeneity in the human population. Understanding of the genetic basis of virulence determinants will provide important insights for antiviral drug and live attenuated vaccine development.

II. METHODS

A. Acquiring of Biological (Sequences) Data

The sequence data were chosen according to the following criteria: availability of the genomic sequences, an estimated basic reproductive number (R_0), and that the IAV subtype can infect human.

Based on these criteria, 96 sequences of IAV were found. The data cover several countries (Boston, Bochum, Honk-Kong, Italy, Johannesburg, Mexico, Netherlands, Sydney, Toronto, Warsaw, and several subtypes, which are H1N1, H3N2, H7N7, and H7N1.

To highlight the local and global similarities among the sequences since our objectives are; among others are elicitation of-conserved regions, detection of simple sequence repeats (SSR), and recognition of motifs (patterns). To do this, raw IAV sequences in fasta-format, served as input for multiple sequence alignment, motifs and repeats by the software T-coffee [11], MEME [12], and MiSA [13] respectively.

B. SSR and Motifs Search

1) Detection of SSRs

To detect SSRs in the EST sequence data sets, we used exact matching algorithms, implemented in a slightly modified version of the Perl script MIcroSAtellite identification tool (MISA). This program take a FASTA formatted sequence file containing multiple sequences, as an input, and search each sequence (contig or singlet) for all possible combination of mono-, di-, tri-, tetra-, and penta- as well as complex repeats with the default criteria of minimum numbers of repeats are set to 9 for di-nucleotides, 6 for tri-nucleotides, 5 for tetra-nucleotides, and 4 for penta-nucleotides.

Two output files are generated by MISA, one file reports the sequence description (including sequence ID and descriptive title), the number of SSR motifs in each sequence, the length and composition of SSR, the number of repeats, the SSR's start and end position, and the total length of the sequence containing the SSR. All identified sequences were then stored in FASTA-format files. Single (mono) nucleotide repeats were not selected because they were generally not considered as useful polymorphic markers.

The results of the MISA runs were transferred to an Excel style worksheet for further analyses.

C. Motif Mining with MEME and Associated Programs

For a set of closely related sequences, often-shared motifs can be discovered using methods based on multiple alignments. Often, distantly related sequences that share common grounds cannot be easily aligned, leading to unsatisfactory results. To detect such subtle patterns, more sophisticated algorithms such as expectation-maximization (EM) are used.

The purpose of MEME (Multiple EM for Motif Elicitation) [14] is to allow users to discover patterns in the DNA or protein quasi-unrelated sequences [12], [15].

1) Rfam

Non-coding RNA genes (ncRNAs) that are not translated into protein, but they may produce as final product functional RNA molecules. Among these functional RNAs are the transfer RNA and ribosomal RNA.

Like the genes encoding proteins, ncRNAs fall into families that have evolved from a common ancestor. By alignments of these gene families ncRNA we can learn about their structure and function. The basic objectives of RFAM RNA database [16] are:

To integrate as many of the existing structural RNA alignments (as well as new alignments) in a common structure annotated format Provide a system for analyzing and automatically annotate sequences (including complete

ISSN 1737-9296

genomic sequences) for detecting the presence of homologous known structural RNA [17].

2) FIMO

FIMO (Find Individual Pattern Occurrences) [18] is a tool that looks for sequences present in a specific database, which share similarities with the patterns obtained by the tool MEME, serving as a template.

The operation of this tool is based on the calculation of the score of the log report (likelihood) of each pattern with the position of the region, and then the score will be converted to p value using dynamic programming methods [19].

D. Blast Search of SSR and Multilevel Motifs

BLASTx (https://blast.ncbi.nlm.nih.gov/Blast.cgi) was performed on translated SSR and multilevel motifs to search proteins with significant match to translated SSR nucleotide sequence. BLASTx was performed against non-redundant (nr) database. For this study, a significant match was defined as a sequence with expected value (E-value) <=1e-3 and identity>=70%.

E. 3D Modeling and Docking

The challenge in developing vaccines against influenza is the ability of the virus to mutate rapidly to evade selective immune pressure. Hemagglutin is the predominant surface glycoprotein and the major determinant of antigenicity. Mutations leading to changes in the HA protein coding sites are often reported. However, genetic sequencing studies predict at best the disruption or creation of new sequence or motifs at this site (HA coding region) or other regions; but they rarely reflect actual phenotypic changes in HA structure, and most importantly the docking energy between HA complex and the sialic receptor. This energy reflects somehow the viral bioactivity. Therefore, combined analysis of evolution (mutation in active site), the distribution of the motifs (found by MEME) repeats in the original IAV sequences and docking (viral bioactivity) to better define the relationships among mutation and motif repeat distribution and actual virulence as reflected by structural change in the HA protein and its docking energy to the human receptor. We combined this information with structural change with binding data to correlate the phenotypic changes with biological activity. To understand the structural basis for site-specific mutation, we performed structural modeling (when needed) and evaluation of docking energy. We investigated the motifs distribution and reduced or increased docking energy.

F. Prediction of 3D Protein Structure:

To model the 3D structure of mutated or original HA proteins we used the approach of modeling by similarity or comparative homology as commonly known. For this purpose the software Modeller [22], [23] is used. Once the 3D model is determined using Modeller, visualization and analysis is made using the software Chimera [22] and Pymol [23]. Models checking are assured by the software for Protein Structure Analysis: Prosa [24], [25] and Rampage [26], [27] for molecular energy and Ramachandran principals.

1) The Receptors

In the docking section, we intended to analyze the interactions between HA sequences and specific receptors. These receptors are located on epithelial cells of the human respiratory tract. The receivers used are of two types (LSTc, and 6'-SLN) in PDB format taken directly from the base RCSB [28]. All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

G. Docking

The ability to attach to and enter the cell is the first phase of the infection, this assured by docking of the HA protein to the cell receptor. To evaluate the quality of this attachment the docking energy of the Crystal, MNRI or computed 3D model protein with the sialic-acid needs to be evaluated, the lower this energy the better the attachment of the virus to the cell receptor

H. Statistical Analysis

A simple statistical linear model (regression test) [29] was used to test for association (correlation) between the R0 and the number of repeats of each SSR type and motifs.

Holm's correction for multiple testing was applied when necessary [30]. Principal components analysis (PCA) [31] was performed, and associated biplot [32] was drawn. We visualized the distribution of the SSR and motifs with the IVA subtype, as reported in the principal components dimensions, energy values were reported as auxiliary variable.

All analysis was performed using the R software for statistical computing version 3.0, and associated packages: FactoMiner and SensoMiner, which are freely available from web [33].

III. RESULTS

Using two criteria; availability of published R_0 in the scientific literature and that the IAV subtype infects humans: 96 sequences have been retrieved with their respective R_0s, covering H1N1, H3N2, H7N7, and H7N1. These sequences in fasta format were used as input for the software T-coffee, MEME and Misa, in order to detect conserved regions, motifs, and SSrs among these sequences.

Alignment, Conserved Motif, and SSR

Global alignment is performed between the segments of the viral genomes, segment by segment. We obtained for each alignment a high score of about 900, this puts into consideration the relationship between subtypes of the virus so these viruses are probably counterparts and they derive from the same common ancestor. However, for the overall multiple alignments with T-coffee, we did not find any conserved alignment-block for most of the sequences through all viral segments.

Using Misa with various parameters, we found 16 complex SSRs of interest. Using regression analysis while adjusting for human population density (Fig.1) only SSR16: (TA)3 was found to be positively associate with R_0, i.e. the number of

ISSN 1737-9296

repeats of SSR16 is correlated with the R_0 values (p-value<0.05).

Using MEME, we were able to detect 4 multilevel motifs of interest. Only three (3) of them: motif-1, motif-2 and motif-3, (their number of repeats) were found to be associated with R_0. While motif-1, and motif2, and, are positively correlated with R_0, motif-3 is negatively correlated with R_0.

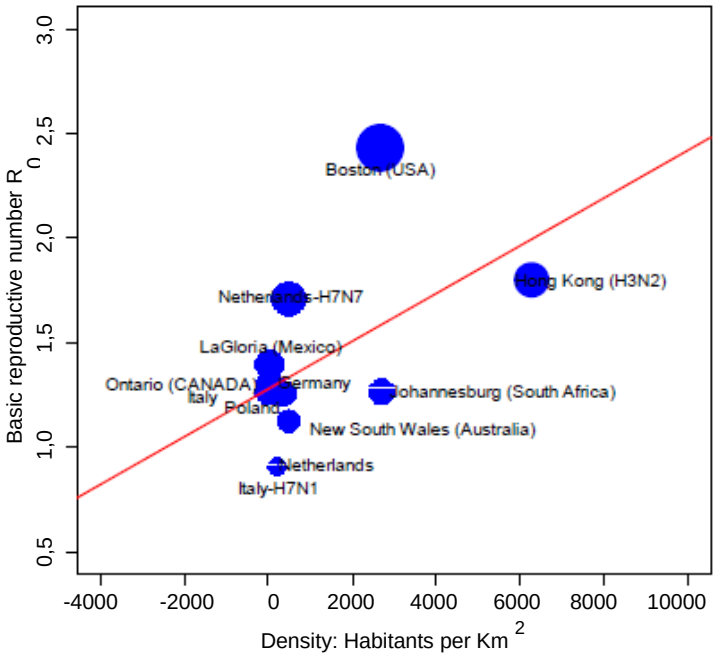


Fig. 1 Regression plot of the basic reproductive number R_0 against the population density in towns and country-wide. The IAV subtype is mentioned when it is different of H1N1.

Searching for Significance and Biological Meaning of the Patterns (motifs) and SSRs

Analysis of patterns and SSRs via Rfam revealed no sequence similarity; this can explained by the small size of the patterns and the SSRs analyzed. To overcome this difficulty, the FIMO tool is used, which plays the same role as Rfam. Each pattern (motif) and each SSR is matched to and positioned on sequences sharing similarity. These sequences are subsequently scanned on the database of "ssRNA negative strand virus" in NCBI. These patterns and SSRs were searched for in other sequences having their PDB format in the PDB database.

Positioning the SSR sand motifs on the DNA sequence retrieved by MAST (MEME suit) shows that these sequences are actually overlapping and they do fall in region coding for functional proteins such as: HA, M1-M2, NA, NP, PA, NS1-NEP (NS2), PB1, PB2. Indeed, Motif -1 overlaps with M1-M2, where matrix protein 1 (M1), is a major structural protein, and the dominant protein in determining virus morphology and also plays an important role in virus assembly and budding [34]. Matrix protein 2 (M2), on the other hand, is the ion channel that regulates the pH, and is responsible for virus uncoating, after virus entry into the cell [35], [36] Motif-1 covers also region coding for NA (neutralizing antibodies), NP, and PA. Motif-2 forms a part of the nucleic sequences, which codes for the proteins M1-M2, NS1-NEP, PA, and PB1. Motif-3 besides coding for M1-M2 and NA, PA, overlaps with region coding for PB1 and PB2. Motif-4 and besides overlapping with region coding for M1-M2, and PA, it codes for PB1. The SSR12 is in the sequence which codes for the protein NP and the SSR16 in the sequence that encodes for the HA (binding and fusion activities), PB2 and PB1 are ribonucleoproteins (RNPs), with NP, PB2 and PB1 are responsible for replication of the viral genome (Table 1).

Thus, the software MEME suite and MiSA allowed us to determine the patterns (common motifs) and SSRs that are involved either in virulence (Fixation or in viral replication).

As multilevel motifs are more comprehensive then SSRs (sequences found by exact matching algorithm), we will be using only multilevel motifs in the subsequent analysis. For all HA protein sequences, with pdb format corresponding to the 11 virus sequences used above (acquired from the NCBI database) (Table 2), a docking analysis with the sialic receptor was performed, and the docking energy evaluated.

The distribution of the 4 motifs and SSRs with respect to these HA proteins was also determined. From the data-table displaying these information and since there are a hefty number of variables, and because their representation is not feasible using the traditional approach, the principal component analysis (PCA) was used. The resulting biplot is drawn (Fig.2a, 2b). Note that motifs 1, and 3 and SSR-10 are in association with docking energy suggesting that these motifs and SSR are involved in the virulence of the virus, through attachment of HA to its sialic receiver. On the other hand, we found that the motif-2, motif-4 and the SSR-12, SSR-16 are independent of the docking energy and therefore they may not be involved in the fixation of the HA on the receptor. These results do not contradict the previously concluded ones (that these regions contribute to virulence). This suggests that motif-2 and SSRs 12 and 16 are rather involved in viral replication. Combining finding of the regression analysis and the principal components analysis shows that motifs and SSRs that are overlapping with genomic regions responsible for viral replication is the most significant. This means that the virus entry in the cell is not the whole story, although it is a necessary and crucial step. Replication and assembly, thus their rate is rather the major determinants of virulence and infectiousness.

ISSN 1737-9296

TABLE I
POSITION OF SSRS AND MOTIFS IN DNA SEQUENCES

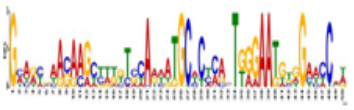
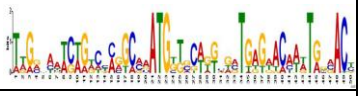
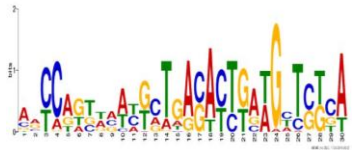
Motif/ SSRs	Protein Encoded & Overlapping SSR	Sequence in Logo-format
Motif1	M1-M2 (SSR4)	
	NA (SSR7/SSR11)	
	NP(SSR3/SSR12/ SSR13)	
	PA (SSR5)	
Motif2	M1-M2(SSR4)	
	NS1-NEP	
	PA (SSR2/SSR5	
	/SSR9/SSR16)	
Motif3	M1-M2 (SSR4)	
	NA (SSR7)	
	PA(SSR5)	
	PB1 (SSR4/SSR8)	
	PB2 (SSR16)	
Motif4	M1-M2 (SSR4)	
	PA (SSR2/SSR5/SSR 9/SSR16)	
	PB1 (SSR4/SSR8)	
SSR16	PB2	(TA)3
	PB1 (SSR4/SSR8)	
	NP(SSR3/SSR12 /SSR13)	
	NA (SSR7/SSR11)	
	HA (SSR10)	

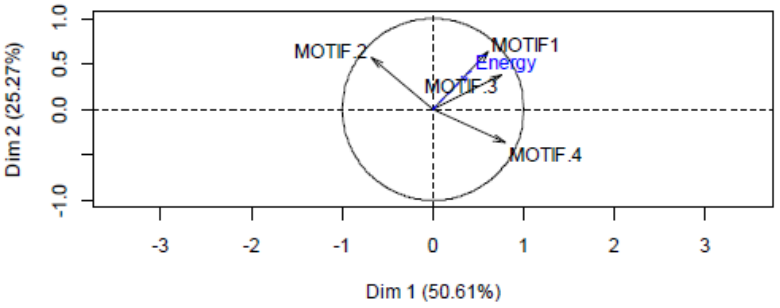
TABLE III
HA SEQUENCES (WILD AND MUTATED), MUTATIONS AND THEIR POSITIONS.

Mutated Sequence	Subt-ype	Wild Sequence	Mutation	Position given "target sequence"
4CR0.pdb	H5N1	3FKU.pdb	N182K	N186K
			G139R	G143R
3KU6.pdb	H2N2	3QQI_A.pdb	Q226L	Q176L
A/Netherlands/33/03	H7N7	4dj6	A143T	A143T
4GXX.pdb	H1N1 (1918)	1RD8.pdb	D190E	D191E
			D225G	D226G

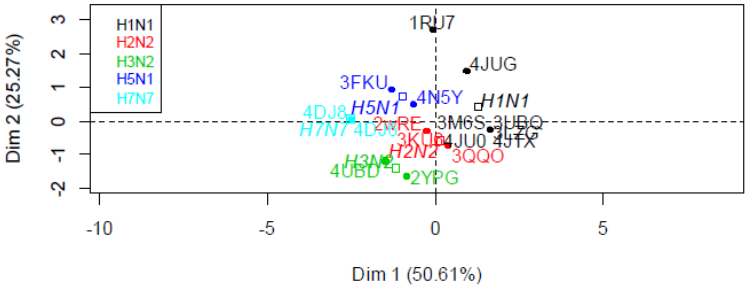
IV. DISCUSSION & CONCLUSION

The viral genome consists of eight molecules of RNA of negative polarity found in the form of ribonucleic complex (RNPv). RNA segments mutations are frequently the cause of epidemics and pandemics because the variability of their genomes makes them evolve very rapidly. This can be achieved by two mechanisms: antigenic drift and re-assortment.

We presented here a new *in silico* approach for the determination of the genetic regions responsible for the virulence of IAV based on R_0 values and whole genetic data of this microorganism. In fact, this method can be applied to other infectious micro-organism with data on R_0 and genome is available. Candidate regions were actually determined using programs like MEME and MISA, however, other bioinformatics tools, for pattern recognition, and sequence repeats that may can be used. The question which program is the best for a particular situation is open, and beyond the scope of this work.



(Fig.2a)Variable Factor Map



(Fig.2b) Individual Factor Map

Fig. 2 Biplot analysis: Motifs and SSR distribution

Two statistical approaches have been used to test for association (correlation) between candidates genetic regions motifs and SSRs found by MEME and MISA and measures quantifying the virulence, which are the R_0 and the docking energy between the HA and the sialic acid.

The first approach consists in regression of the number of repeats against R_0, which yields to the identification of motif-1, motif2, motif3, and SSR16, these regions overlap with those coding for the proteins: M1-M2, NS1-NS2, PA, PB1, and PB2. In fact, M1 is major surface component of the virus and it is responsible for virus budding, nuclear export and assembly [34]. M2 is an ion channel and a ph regulator for HA synthesis [36]. NS1-NS2 are responsible for control of gene expression and export of RNP from the nucleus during viral replication. Moreover, the polymerase complex responsible for viral transcription and replication is formed by PB1, PB2, and PA [37]. Most of these proteins are responsible for virus replication. Residues in the protein PA have identified as contributing to H7N3 virus virulence when infecting mice [38].

PB1-F2 protein is a factor of virulence of influenza A viruses which increases the mortality and morbidity associated

ISSN 1737-9296

with infection [39]. Most seasonal H1N1 Influenza A viruses express a truncated form of PB1-F2 [40].

The second approach consists in the study of the distribution of the motifs and SSRs with the docking energy-of HA and human receptor, this led to association between motif1, and motif3, which code for M1-M2, NA, and PA. SSR16 is also associated with the docking energy, and is situated in part of the region coding for HA.

The results from the first approach seems to more consistent as the ability and rate of the virus replication is more determinant to the virulence. The ability of docking and entry is also crucial but the virus either gets in the cell or not. While, the severity of influenza infection is not only influenced by viral virulence factors but also by individual differences in the host response, we adjusted in this analysis only for the density of human population, adjusting for the human genetic background is very difficult and requires identification of specific genetic region, and may even requires controlled experimentations.

ACKNOWLEDGMENT

This study received financial support from the Secretariat of the State for Scientific Research, Technology, and Competencies Development in Tunisia, through funding of Research Program Contract for Institute Pasteur of Tunis (2011-2015).

REFERENCES

[1] W. D. Tanner, D. J. Toth, A.V. Gundlapalli, "The pandemic potential of avian influenza A(H7N9) virus: a review," *Epidemiol Infect*, vol.143(16), pp. 3359-3374, Dec. 2015.

[2] T. Chen et R. Zhang, "Symptoms seem to be mild in children infected with avian influenza A (H5N6) and other subtypes," *J Infect*, vol. 71(6), pp.702-703, 2015.

[3] S. J. Gamblin, J. J. Skehel, "Influenza hemagglutinin and neuraminidase membrane glycoproteins," *J Biol Chem*, vol. 285(37), pp. 28403-28409, Sep. 2010.

[4] T. Bedford, S. Cobey, P. Beerli, M. Pascual, "Global migration dynamics underlie evolution and persistence of human influenza A (H3N2)," *Plos Pathog*, vol. 6(5), 2010.

[5] P. Wright G. Neumann, Y. Kawaoka. Orthomyxoviruses. *In Fields Virology* 2013. 6th Edition, Knipe DM et Howley P : Lippincott Williams & Wilkins; Chapter: 411186-411243.

[6] X. Feng, Z. Wang, J. Shi, G. Deng, H. Kong, S. Tao, C. Li, L. Liu, Y. Guan, H. Chen, "Glycine at Position 622 in PB1 Contributes to the Virulence of H5N1 Avian Influenza Virus in Mice," *J Virol*, vol. 90 (4), pp. 1872-1879, Dec. 9, 2015.

[7] H. Li, B.Cao, "Pandemic and Avian Influenza A Viruses in Humans: Epidemiology, Virology, Clinical Characteristics, and Treatment Strategy," *Clin Chest Med*, vol. 38(1), pp. 3859-3870, Mar. 2017.

[8] K. K. To, J. F. Chan, K. Y. Yuen, " Viral lung infections: epidemiology, virology, clinical features, and management of avian influenza A(H7N9)," *Curr Opin Pulm Me*, vol. 20(3), pp. 225-232, May. 2014.

[9] K. J. Taubenberger and D. M. Morens, "The Pathology of Influenza Virus Infections," *PMC Annu Rev Pathol*, vol. 3, pp. 499-522, 2008.

[10] S. Heffernam, Smith, J. R. Wahl, "Perspectives on the basic reproductive ratio,"*Soc Interface*, vol. 2(4), pp. 281-293, Sep. 22, 2005

[11] C. Notredame, D.G. Higgins, " T-Coffee: A novel method for fast and accurate multiple sequence alignment,"*J Mol Biol*, vol. 302(1), pp. 205-217, Sep. 8, 2000.

[12] L. Bailey, J. Johnson, G.E. CGrant, and W. S. Noble, " The MEME Suite," *Nucleic Acids Res*, vol 43, w39-w49, Jul. 2015.

[13] T. Thiel, W. Michalek, R. K. Varshney, A. Graner, "Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare L.*),"*Theor Appl Gene*, vol. 106(3), pp. 411-422, 2003.

[14] The MEME website. [Online]. Available:http://meme.nbcr.net.

[15] L. T. Bailey, N. Williams, C. Misleh, and W. Li, "MEME: discovering and analyzing DNA and protein sequence motifs," *Nucleic Acids Res*, vol. 34, w369-w373, Jul. 2006.

[16] The rfam website. [Online]. Available: http://rfam.xfam.org/

[17] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, R. Eddy, "Rfam: an RNA family database,"*Nucleic Acids Res*, vol.31(1), pp 439-441, Jan. 1, 2003 .

[18] The FIMO website. [Online].Available http://meme-suite.org/tools/fimo

[19] C. E. Grant, T. L. Bailey, W. S. Noble, "FIMO: scanning for occurrences of a given motif,"*Bioinformatics*, vol. 27(7), pp. 1017-1018, Apr. 1, 2011.

[20] The modeller website. [Online]. Available: https://salilab.org/modeller; version 9.15.

[21] B. Webb, A. Sali, " Protein Structure Modeling with MODELLER," *Methods Mol Biol*, vol.1654, pp. 39-54, 2017

[22] The chimera website. [Online]. Available: https://www.cgl.ucsf.edu/chimera/; version v1.10.2.

[23] N. Matthews, R. Easdon, A. Kitao, S. Hayward, S. Laycock, "High quality rendering of protein dynamics in space filling mode", *J Mol Graph Model*, vol. 78, pp. 158-167, Nov. 2017.

[24] The prosa website. [Online]. Available: https://prosa.services.came.sbg.ac.at; web version.

[25] F. Nazmi, M. A. Moosavi, M. Rahmati, Hoessinpour-Feizi MA4. "Modeling and structural analysis of human Guanine nucleotide-binding protein-like 3, nucleostemin," *Bioinformation*, vol. 11(7), pp. 353-358, Jul. 31, 2015.

[26] The rampage website. [Online]. Available: http://mordred.bioc.cam.ac.uk/~rapper/rampage.php

[27] M .Wiederstein and J. Sippl, "ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins," *Nucleic Acids Res*, vol. 35(2), pp. 407-410, Jul. 2007.

[28] The RCSB website. [Online]. Available: http://www.rcsb.org/pdb/home/home.do

[29] J. M. Chambers, *Linear models*, Chapter 4 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole, *J Pacif Gro*, 1992.

[30] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, pp. 65-70, 1979.

[31] W. N. Venables, and B. D. Ripley. *Modern Applied Statistics with S*, Springer-Verlag, 2002.

[32] K. R. Gabriel. "The biplot graphical display of matrices with application to principal component analysis," *Biometrika*, vol. 58, pp. 453-467, 1971.

[33] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available: http://www.R-project.org. 2008.

[34] J. S. Rossman, R. A. Lamb. "Influenza virus assembly and budding", *Virology*, vol. 411(2), pp. 229-36, Mar. 15, 2011.

[35] A. Helenius, "Unpacking the incoming influenza virus", *Cell*, vol. 69(4), pp. 577-578, May. 15, 1992 .

[36] L .J. Holsinger, D. Nichani , L. H. Pinto , R. A. Lamb," Influenza A virus M2 ion channel protein: a structure-function analysis", *J Virol*, vol. 68(3), pp. 1551-1563, Mar. 1994 .

[37] B. W. Jagger , H. M. Wise, J. C. Kash , K. A. Walters , N. M. Wills, Y. L. Xiao , R.L. Dunfee , L. M. Schwartzman , A. Ozinsky , G. L. Bell et al , "An overlapping protein-coding region in influenza A virus segment 3 modulates the host response," *Science*, vol. 337(6091), pp.199-204, Jul. 13, 2012.

[38] B. L. DesRochers BL, R.E. Chen, A.P. Gounder, A. K. Pinto, T. Bricker, C. N. Linton, C. D. Rogers, G. D. Williams, R. J. Webby, A. C. Boon, "Residues in the PB2 and PA genes contribute to the pathogenicity of avian H7N3 influenza A virus in DBA/2 mice," *Virology*, vol. 494, pp. 89-99, Jul. 2016.

[39] H. Sediri, S. Thiele, F. Schwalm, G. Gabriel, H. D. Klenk, "PB2 subunit of avian influenza virus subtype H9N2: a pandemic risk factor," *J Gen Virol* , vol. 97(1), pp. 39-48, Jan. 2016 .

ISSN 1737-9296

[40] D. Ajjaji, C. A. Richard, S. Mazerat, C. Chevalier, J. Vidic, "N-terminal domain of PB1-F2 protein of influenza A virus can fold into amyloid-like oligomers and damage cholesterol and cardiolipid containing membranes," *Biochem Biophys Res Commun*, vol. 477(1), pp.27-32, Aug. 12, 2016.

ISSN 1737-9296