Analysis of Influenza Virus Proteins Based on its Primary and Secondary Structure

P. Manikandan[1], D. Ramyachitra[2], C. Muthu[3]*
[1,3]Department of Data Science, Loyola College, Chennai-600 034, India
[2]Department of Computer Science, Bharathiar University, Coimbatore-641 046, India.

Abstract

An infectious disease named swine flu is triggered by an influenza virus. The Influenza virus is any strain of influenza family virus that is mutual in pigs. Mainly, the subtypes of influenza virus H1N1 in the swine classification of Influenza A, affects the humans. For identifying the function of proteins, knowing the structures of the proteins is an important task. In this research work, the secondary structure of Influenza virus protein, predicted by the SOPM (Self Optimized Prediction Method) server is given as input to meet the objectives such as, (i) Individual Amino acid count for every swine flu virus protein, (ii) Predicting the Amino acid patterns (iii) Similarity analysis of predicted patterns based on alpha helix secondary structure. In Similarity analysis, two types of analysis are performed such as finding the exact and most similar protein patterns. The results show that the homologous sequences are predicted and similarity between the protein patterns are compared effectively. And also, the experimental results reveal that the properties of the sequence such as Hydrophobic and Hydrophilic are identified in an efficient way.

Keywords: *Influenza A Virus*, *Homosapiens*, *Mus Musculus*, Swine Flu, Protein Secondary structure, Amino Acid Patterns.

1. Introduction

Swine flu is one of the recent killer diseases spreading in the modern world. Mainly the pigs are affected by this disease, which are normally diverse from the human viruses. The symptoms which are most reported for the swine flu disease is cough, fever, sore throat and myalgia. The swine Influenza A (H1N1) virus spreads fast through human to human transmission in several countries. For the respiratory region infections, influenza viruses play a vital role and it is responsible for 3-5 million clinical infections and 250,000-500,000 fatal cases per annum (Stohr, 2002). Generally, the transmission of swine flu in human infections with different viruses occurs in people with contact to infected pigs. The transmission occurs by consumption of diseased droplets and by direct interaction, which is aid by land travel, air and social gatherings (Manish Sinha, 2009). Also, the transmission of swine flu in human can be possible by eating the pork or other products resulting from pigs. The swine flu has several limitations such as limited treatment, regular need for individuals, high risk for secondary infection are critical to control of swine influenza out breaks (Rewar, S, et al., 2015). Mainly swine flu disease affects several organisms in the real world that includes *Homosapiens*, *Influenza A Virus*, *Mus Musculus*, Streptomyces aureofaciens and other organisms. To overcome

this crisis, this research work focused on swine flu disease in three different organisms such as *Homosapiens*, *Influenza A Virus* and *Mus Musculus* for identifying the similarity and differences of the patterns between the protein sequences.

In protein secondary structure, the alpha helix is a spiral conformation in every backbone N-H groups. The alpha helix motif consists of a group of three or four residues along with the protein sequence (Dunitz, 2001). The alpha helix is the major predictable structure from the local structure in protein sequences (https://en.wikipedia.org/wiki/Alpha_helix). Every protein consists of 30% alpha helix structure in nature and some amino acids occur more often than the others known as helix propensity (C.Nick Pace and J.Martin Scholtz, 1998). In general, the proteins which consist of more alpha helixes also have more molecular weight and the structure does not change in nature. Several methods can be used to analyze the protein secondary structure such as YASPIN, PredictProtein, 1D Protein Structure Prediction Server, Jpred and so on. In YASPIN (Kuang Lin, et al., 2004) method, the protein structure can be predicted by using the HNN (Hidden Neural Network) and also it uses the PSI-BLAST algorithm to produce the PSSM (Position-Specific Scoring Matrix). The PredictProtein method achieves the protein structure analysis by significantly expanding the size of structural annotations (Guy Yachdav et al., 2014).

1D Protein Structure Prediction Server uses a multiple linear regression model and feature based sequence representation to predict the helix and strands from the sequences (Leila Homaeian, et al., 2007). In Jpred method, the structure analysis can be analyzed by using the JNet algorithm (Alexy Drozdetskiy et al., 2015). The Proteins which are having related functions may not illustrate high homology however may contain sequences of amino acid residues that are highly conserved (Nicholas Hulo, et al., 2006). Applications of structure similarity includes the matching of protein structures, protein classification, finding the protein motif in protein structure and finding the pharmocaphore (a substructure common to all the ligands) in ligands. In this research the protein structure similarity analysis is carried out by using the amino acid patterns corresponding to the protein structural classes that can be predicted by the SOPM tool (Geourjon C and Deléage G, 1995).

To identify the motif pattern and homology sequences, this research work focused on the protein structure similarity analysis. Based on the literatures, this research focus on the alpha helix secondary structure and considers three or more residues as alpha helix patterns. By using this way of analysis, the wetlab researchers conclude that the alpha helix will occur, only when the particular amino acid pattern likely to be similar to the patterns in the alpha helix database. The remaining section of the paper is organized as follows. Section 2 describes the methodology and Section 3 shows the experiments on the Influenza protein sequence and discusses the results. Finally, the conclusion and future enhancement is given in Section 4.

2. Methodology

In this research work the alpha helix secondary structure pattern is analyzed by using the Influenza protein structural data [Secondary data] based on its primary sequence [Fasta]. The secondary structural data is predicted and classified as patterns using the java environment. For analyzing the influenza virus patterns Knuth–Morris–Pratt (KMP) algorithm is used.

2.1 Datasets

Protein Data Bank (PDB) is a structural information repository for 3 dimensional structures of proteins and it has been used as benchmark for this study. The benchmark datasets are collected from PDB database, and this research work mainly focuses on swine flu disease. This research work spotlight on three specific organisms namely *Homosapiens*, *Influenza A Virus* and *Mus Musculus*. For the analysis 110 sequences have been taken in *Homosapiens*, 255 sequences in *Influenza A Virus* and 50 sequences in *Mus Musculus*. Most of the *Homosapiens* and *Mus Musculus* sequences are tested on X-Ray method only, but the *Influenza A Virus* sequences are tested on different method such as X-Ray, Solution NMR and Electron Microscopy.

2.2 Amino Acid Count

In this step, the most occurrence of amino acid is predicted to find out the properties such as hydrophilic and hydrophobic of particular protein. In this research, the most occurrence of amino acid protein is predicted and it can be used for wetlab researchers for analysis of particular protein. This research work predicts the amino acid count based on the alpha helix secondary structure of every protein.

2.3 Amino Acid Pattern

In this part, the amino acid patterns are predicted based on the alpha helix secondary structure of every protein. For this experimental analysis, a database has been created for storing the patterns of amino acid based on alpha helix secondary structure. In general, three bases of code represent a structure. Hence, the patterns which consists of more than or equal to three amino acid codes considers as pattern. Based on the amino acid patterns from Tables 1-5, the wetlab researchers can analyzed whether that protein belongs to the alpha helix structure or not.

2.4 Similarity Analysis

This part of work is used to analyze the protein patterns based on the alpha helix secondary structure of every protein. Two types of protein pattern analysis have been carried out in this research work. They are as follows.

1. Exact similar protein patterns
2. Most similar protein patterns

In the first phase, the exact proteins are identified for removing the repeated protein chains. The identification of repeated sequences will be more helpful to researchers working in wetlab field for analyzing the proteins. And in the second phase, the proteins which are similar in alpha helix patterns are identified, for the reason that the protein which consists of more alpha helix will be more stable in nature. Also based on the similarity analysis of proteins, the functional and evolutionary relationships between the proteins are identified in an easy manner by using this research work. For analyzing the exact and most similar protein patterns of the influenza virus, the Knuth–Morris–Pratt (KMP) algorithm (D.E. Knuth, et al., 1977) is used.

3. Experimental Results

This research work focuses on the similarity pattern analysis of the influenza virus based on alpha helix secondary structure. The Supplementary Table 1-4 shows the amino acid count for 3MGT, 1WBY, 2CII and 1HSB proteins based on alpha helix secondary structure. For every protein and its chain, the amino acid count will be predicted. The Hydrophobic Amino Acids are Glycine (G), Alanine (A), Valine (V), Leucine (L), Isoleucine (I), Phenylalanine (F), Proline (P) and the Hydrophilic Amino Acids are Asparagine (N), Cysteine (C), Lysine (K), Aspartic (D), Methionine (M), Threonine (T), Glutamine (Q), Histidine (H), Arginine (R), Glutamic (E), Tryptophan (W), Tyrosine (Y), Serine (S). Based on the count of individual amino acid, the wetlab researchers can analyze that which amino acid has majority in that particular protein and also, they make decision on that protein whether it belongs to hydrophobic or hydrophilic amino acid. Two types of protein pattern analysis have been carried out in this research work. The first type of analysis is to finding the exact similar protein patterns for removing the repeated protein sequences. And in second type of analysis, the most similar protein patterns are identified. Table 1 shows the exact similar proteins pattern and details for the *Homosapiens* and Table 2 shows the protein patterns and details for the Infuenza A Virus. There is no exact similar protein in *Mus Musculus*.

Table 1: Exact similar proteins for *Homosapiens* Organism

S.No	Protein Id	Patterns	Mutation
1.	2FOZ 2F10 2F11 2F12 2F13	SLLAFAEQRA HAE WQAQEVVAQARL EQQQLQTRANVT DAA AYREW ARS AYAY QDTLECQVAEV HLRAR FQESQLVKKL RADLGAY SDLQS YEEIVFLMFTLK	No Mutation
2.	3G81 3G83 3G84	AMADIGSDVASLRQQVEALQGQVQHLQAAFSQYKK VGEK EAQ SAAENAALQQLVVA EAA	- - yes

3.	2X4N 2X4P 2X4Q 2X4S 2X70 1HHI 1HHK 2VLJ 2VLK	AASQ TRK TLRG SHTVQRMY WRFLR LKEDLRSWTAADMAAQTTKHKWEAAHVAEQLRAYLE CVEWLRRYLE ETL HEATLRCW	M SN EYACR	MQLL VFSSLQWY VKKLKR MRSM FACANAF	HDA YSQIVNDF IA Y TEDLK AEISHTQKA LSSRL VTQ SAEA	No Mutation

The Table 1 shows the exact similar patterns for the *Homosapiens* organism. For every pattern shown in the Table 1, the user can choose any one of the proteins rather than choosing all the protein sequences for every pattern. For example, in Table 1 the protein ID's such as 2FOZ, 2F10, 2F11, 2F12 and 2F13 (S.No: 1) are having the exact similar alpha helix pattern. Instead of testing all the protein sequences, the wetlab researchers can choose any one of the protein sequences. Similarly, for the second and third pattern shown in Table 1, the user can choose any one of the protein sequences.

Table 2: Exact similar proteins for *Influenza A Virus* Organism

S.No	Protein Id	Patterns		Mutation
1.	1VOZ 1W1X 1W20 1W21 2CML	FRA KIEELA AQHIEECSC MMTHTSKYLCS YWA		No Mutation
2.	1ING 1INH 1INW 1INX	VEY QHVEEC KDL		Yes Yes Yes Yes
3.	1BJI 1L7F 1L7G 1L7H 1MWE 1NNA 1NNB 1XOE 1XOG 2C4A 2C4L	FNNLTK NTWARNILRTQ KHIEECSC DGVNTW FMDYWA CYRAC TEFL		- - Yes Yes - - - - - - -
4.	1MQL 1MQM 1MQN	TNATELVQ ACTLIDALL RSN YASLRSLVAS SRLNWL KQNTLKLATG	LFGAIAGFIE AADLKSTQAAIDQINRKLNRVIEKTNEKFHQIEKEFS EVEGRIQDLEKYVED WSYNAEL ALE LADSEMNKLFEKTRRQLRENAEDM CIESI RDEALNNR FAISCLLLCVVLLGFIMWAC	No Mutation

The Table 2 shows the exact similar patterns for the *Influenza A Virus*. Similarly, for every pattern shown in the Table 2 the user can choose any one of the proteins rather than choosing all the protein sequences for every pattern. For example, in Table 2 the protein ID's such as 1VOZ, 1W1X, 1W20, 1W21 and 2CML (S.No:1) are having the exact similar alpha helix pattern. Instead of testing all the protein sequences, the wetlab researchers can choose any one of the

protein sequences. Similarly, for the patterns 2, 3 and 4 shown in Table 2, the user can choose any one of the protein sequences.

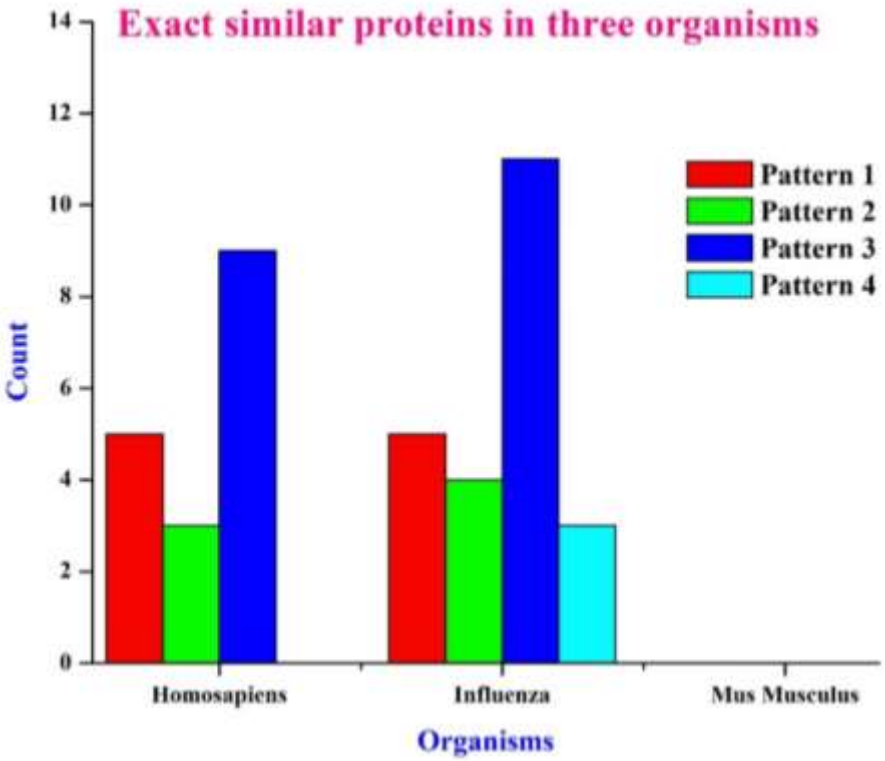

Fig. 1: Total number of exact similar proteins in three organisms

From the Table 1 it is inferred that the *Homosapiens* consists of three types of patterns and the protein id's shown in Table 1 for every protein pattern are exactly same. The *Influenza A Virus* consists of four patterns and the protein id's shown in Table 2 for every protein pattern are exactly same. The *Mus Musculus* does not have any exact similar protein patterns. The total number of exact similar protein patterns for three organisms is shown in Fig.1. The more similar protein patterns and its mutation details of three organisms are shown in the Table 3-5. From these analyses, the wetlab researchers can easily identify the more similar proteins in the three organisms for the swine flu disease. Based on the patterns the drugs can be designed for the swine flu disease.

Table 3: Most similar proteins for *Homosapiens* Organism

S.No	Protein Id	Patterns	Mutation
1.	1DLH	KET	-
	1AQD	LEE FASFEAQGALANIAVDKANLE	-
	1FYT	VYDC	yes
	1HXY	LLKH	-
	1KG0		-

	1H15		-
	1PYW		yes
	1R5I		-
	2G9H		-
	2IPK		yes
	2ICW		yes
	2WBJ		-
	2XN9		yes
2.	1DLH	RVRLLERCIY AEYWNSQKDLLEQRRAAVDTYC	-
	1KG0		-
	1FYT		yes
	1HXY		-
	1LO5		yes
	1PYW		yes
	1R5I		-
	2G9H		-
	2ICW		yes
	2IPK		yes
	2XN9		yes
	1AQD		-
	1H15		-
3.	2VLJ	MQLL VFSSLQWY VKKLKR MRSM FACANAF	-
	2VLK		-
	2VLR		-
	2XN9		yes
	2WBJ		-
4.	1HHI	EYACR	No Mutation
	1HHK		
	2BST		
	2VLJ		
	2VLK		
	2VLL		
	2VLR		
	2X4N		
	2X4P		

	2X4Q 2X4S 2X70 3LKN 3LKO 3LKP 3LKQ 3LKR 3LKS 3MGO 1HSB 2CII		
5.	3FKQ 3GBM	TEQVDTIME AQDILEK YEELKHLLSRINH HEAS SFFRN AEQTK TLN SELEY RKKR	No Mutation
6.	3FKQ 3GBM	LFGAIAGF KESTQKAIDGVTNKVNSIIDKMNTQFEAVGREFNNLERRI ENLNKKME TYNAEL DSNVKNLYDKVRLQLRDNAKEL NECMESVR EEARLKREEISS	No Mutation
7.	1HHI 1HHK 1HSB 2VLJ 2VLK 2VLR 2VLL 2X4N 2X4P 2X4Q 2X4S	AASQ TRK TLRG SHTVQRMY WRFLR LKEDLRSWTAADMAAQTTKHKWEAAHVAEQLRAYLE CVEWLRRYLE HEATLRCW	No Mutation

	2X70		
	3MGO		
8.	2VLJ	HDA	-
	2VLK	YSQIVNDF TEDLK	-
	2VLR	AEISHTQKA	-
	2XN9	LSSRL VTQ	yes
	2WBJ	SAEA	-

The Table 3 shows the most similar patterns for the *Homosapiens*. From this table, the user can identify the most similar protein sequences patterns for the *Homosapiens*.

Table 4: Most similar proteins for *Influenza A Virus* Organism

S.No	Protein Id	Patterns	Mutation
1.	1INY	FNNLTK	Yes
	2B8H	NTWARNILRTQ	Yes
	1A14	KHIEECSC	-
	1L7F	DGVNTW	-
	1BJI	FMDYWA	-
	1L7G	CYRAC	Yes
	1L7H	TEFL	Yes
	1MWE		-
	1NMC		-
	1NNA		-
	1NNB		-
	1XOE		-
	1XOG		-
	2C4A		-
	2C4L		-
2.	1A14	QK	-
	1NMC	SS	-
	1NMA		Yes
	1NMB		yes
3.	1NMC	LTQTTSSLSAS	-
	1NMA	ISNYLNWY	Yes
	1NMB	QEDIA	yes
4.	1RD8	TVDTVLE	-
	1RV0	LLE YEELREQLSSVSS VTAACSYA SSFYRNLL SKS QSLYQNADA QAGRMN YVRSTKLRMA	-
5.	1RD8	LFGAIAG	-

	1RV0	QNAIDGITNKVNSVIEKMNTQFTAVGKEFNNLERRIENLNKKV TYNAEL DSNVRNLYEKVKSQLKNNAKEI DACMESVR EES REEIDGVRSLV	-

The Table 4 shows the most similar patterns for the *Influenza A Virus*. From this table the user can identify the most similar protein sequences for *Influenza A Virus*.

Table 5: Most similar proteins for *Mus Musculus* Organism

S.No	Protein Id	Patterns	Mutation
1.	1NMB	ATA	No Mutation
	1NCD	ASQTIINNY EERASREFNNLTKGL NTWARNILRTQESECV HIEECSC FMDYWA CYRAC TEFL YFL	
2.	2VIT	VTNATELVQ	yes
	2VIR	CTLIDALL	-
	2VIS	ERSKA YASLRSLVAS SRLN	yes
3.	3CPL	MRYFET	-
	1HOC	WME	-
	1WBX	EYWERETQKA	-
	1WBY	QEQW	-
	1YN6	LRNLLGYY	-
	1YN7	TLQQM	yes
	2VE6	DWRLLRGYL	-
	3BUY	LNEDLKTWTAADMAA	-
	3CPL	AAEHYKAYLEG	-
	3FTG	LKN	-
	3PQY	WALGFYPAD	-
	4HUU	WQLNGEELTQDMELVETRPAG	-
	4HUV		-
	4HUW		-
	4HUX		-
	4HV8		-
	4L8C		-
	4L8B		-
4.	1NCB	EEF	-
	1NCD	TAN INN AVLQ	-
5.	1NMA	QK	-
	1NMB	SS	-

6.	1NCA	FNNLTK	-
	1NCB	NTWARNILRTQ	-
	1NMC	KHIEECSC	-
	1NCC	DGVNTWLGRTISRA FMDYWA CYRAC	-
7.	1NCD	IREFNNLTK	-
	1NMA	NTWARNILRTQESECV HIEECSC FMDYWA CYRAC TEFL YFL	-
8.	1HOC	MRYFET	-
	1WBX	WME	-
	1WBY	EYWERETQKA	-
	1YN6	QEQW	-
	1YN7	LRNLLGYY	yes
	2VE6	TLQQM	-
	3BUY	DWRLLRGYL	-
	3CPL	LNEDLKTWTAADMAAQITRRKW	-
	3FTG	AAEHYKAYLEGECVEWLHRYL	-
	3PQY	EELTQDMELVE	-
	4HUU	QKW	-
	4HUV		-
	4HUW		-
	4HUX		-
	4HV8		-
	4L8C		-
	4L8B		-
	3CPL		-
	2CII		-
9.	2CLV	RARWME	-
	2CLZ	EYWERETQKA	-
	1WBZ	LRTLLGY LNEDLKTWTAADMAALITKHWEQAGEAERLRAYLEGTCVEWLRRYLK EELIQDMELV	-
10.	4GMS	EDLAEYFCQQ	-
	4GMT	EQLT	-
	4F15	FLN KDEYERH	-

The Table 5 shows the most similar patterns for the *Mus Musculus*. From this table, the user can identify the most similar protein sequences for the *Mus Musculus*.

3.1. Discussion on Mutation Analysis

In support of the exact similar proteins for the *Homosapiens* in Table 1 the Protein id 3G84, and for the *Influenza A Virus* organism in Table 2 the protein id's such as 1ING, 1INH, 1INW, 1INX, 1L7G and 1L7H are mutated proteins. But the mutation cannot happen in the alpha helix

structure and hence conclude that the mutation does not make any structural changes in alpha helix. With respect to the most similar protein patterns, the mutation has occurred in the protein ids such as 2IPK, 2ICW, 1PYW, 1LO5, 1FYT and 2XN9 for the *Homosapiens* in Table 3 and the mutation does not affect the alpha helix structure. And also, the mutation has occurred in the protein ids such as 1INY, 2B8H, 1L7G, 1L7H, 1NMA and 1NMB for the *Influenza A Virus* from Table 4, and the mutation does not affect the alpha helix structure. For the *Mus Musculus*, the mutation has occurred in the protein ids such as 2VIT, 2VIS and 1YN7 from Table 5, and the mutation does not affect the alpha helix structure.

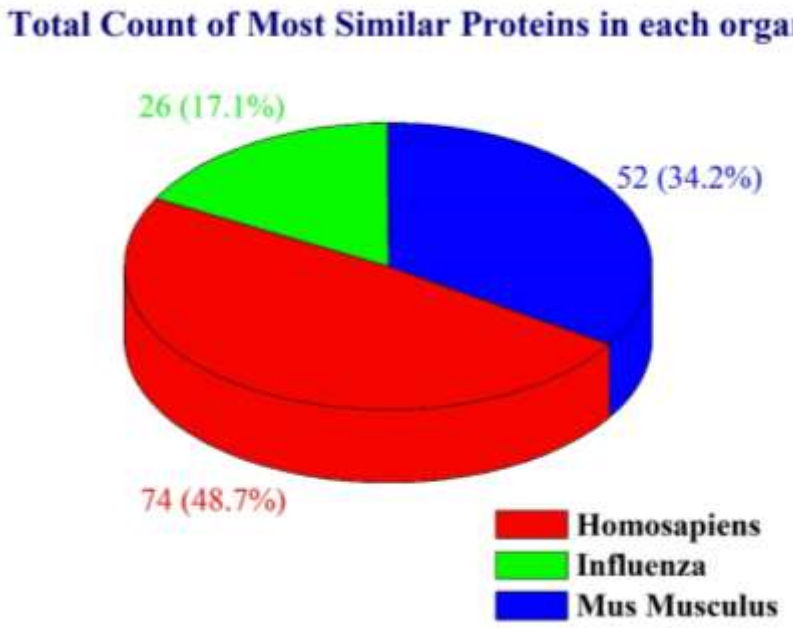

Fig. 2: Total number of most similar proteins in three organisms

3.2. Discussion on Relationship between organisms based on Amino Acid Patterns

Based on the above tables, it is observed that the pattern files in Tables 4 and 5 have similar patterns between them. The Table 4 belongs to the *Influenza A Virus* and the Table 5 belongs to *Mus Musculus*. From the analysis it is inferred that *Influenza A Virus* and *Mus Musculus* are related in swine flu disease. And also, it is identified that the pattern tables such as 3 and 4 have similar relationship with each other and it belongs to *Homosapiens* and *Influenza A Virus* respectively. Hence this research work recognizes that there will be relationship between the *Homosapiens* and *Influenza A Virus*. In *Mus Musculus*, the pattern shown in Table 5 have relationships with each other patterns.

The Tables 1-5 shows the pattern analysis on swine flu disease with respect to alpha helix secondary structure. And the Fig. 2 shows the total number of most similar proteins for every protein patterns. From the results the researchers can identify the alpha helix structure. The Tables 1-5 have proven that, if those patterns occur the alpha helix structure will occur.

4. Conclusion and Future Enhancement

Prediction of alpha helix pattern in a protein sequence is an interesting area in computational biology. In this research work, the exact and most similar protein patterns are identified by using the pattern of alpha helix structure. By identifying the exact similar proteins, the researcher can use any one of the protein sequences for their research, rather than testing all the exact similar proteins. For the wetlab researchers, this research work provides a huge support for identifying the similar proteins in fewer time. Based on this research work the researcher can conclude the protein function similarity and structural stability between the proteins. And also, the functional changes in protein sequence can also be identified. In this study a limited number of datasets have been chosen for analyzing the protein patterns. A large number of datasets have been taken for predicting the alpha helix pattern in unknown raw sequences. In future this research work can be extended to overcome these limitations. And also, this research work can be extended to predict the other protein secondary structures such as Beta sheet, Turn and Coil.

References

1. Stohr, K., 2002. Influenza – WHO cares. Lancet Infect. Dis. 2 (9), 517.
2. http://www.who.int/influenza/surveillance_monitoring/updates/latest_update_GIP_surveillance/en/ (accessed 28.06.2016)
3. Manish Sinha, “Swine flu”, Journal of Infection and Public Health, Volume 2, Issue 4, 2009, Pages 157–166, doi:10.1016/j.jiph.2009.08.006
4. "Key Facts about Swine Influenza (Swine Flu)". *Centers for Disease Control and Prevention* (http://www.cdc.gov/flu/swineflu/keyfacts-variant.htm)
5. Dunitz, J (2001). "Pauling's Left-Handed α-Helix". Angewandte Chemie International Edition. 40 (22): 4167–4173. doi:10.1002/1521-3773(20011119)40:22<4167::AID-ANIE4167>3.0.CO;2-Q.
6. https://en.wikipedia.org/wiki/Alpha_helix
7. C.Nick Pace and J.Martin Scholtz, “A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins”, Biophysical Journal, Volume 75, Issue 1, July 1998, Pages 422–427, DOI: http://dx.doi.org/10.1016/S0006-3495(98)77529-0
8. D.E. Knuth, J.H. Morris, V.R. Pratt, “Fast pattern matching in strings”, SIAM J. Comput., 6 (1977), pp. 323‑350
9. Lin K, et al., “A simple and fast secondary structure prediction method using hidden neural networks”, Bioinformatics. 2005 Jan 15;21(2):152-9. Epub 2004 Sep 17.
10. Guy Yachdav, et al., “PredictProtein—an open resource for online prediction of protein structural and functional features”, Nucleic Acids Res. 2014 Jul 1; 42(Web Server issue): W337–W343, doi: 10.1093/nar/gku366
11. Homaeian L, et al., “Prediction of protein secondary structure content for the twilight zone sequences”, Proteins. 2007 Nov 15;69(3):486-98.
12. Alexey Drozdetskiy, et al., “JPred4: a protein secondary structure prediction server”, Nucl. Acids Res. (2015), doi: 10.1093/nar/gkv332.
13. Geourjon C and Deléage G. “SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments”, Comput Appl Biosci. 1995 Dec;11(6):681-4.
14. Nicholas Hulo, et al., “The PROSITE database”, Nucleic Acids Research, 2006, Vol. 34, Database issue D227–D230 doi:10.1093/nar/gkj063.
15. Rewar, S., Mirdha, D., & Rewar, P. (2015). Treatment and Prevention of Pandemic H1N1 Influenza, Annals of Global Health, 81(5). http://doi.org/10.1016/j.aogh.2015.08.014

Supplementary

Table 1: Amino acid count for 3MGT-H

Amino acid	3MGT:A	3MGT:B	3MGT:C	3MGT:D	3MGT:E	3MGT:F	3MGT:G	3MGT:H	3MGT:I	3MGT:J	3MGT:K	3MGT:L
A	11	1	0	11	1	0	11	1	0	11	1	0
C	2	1	0	2	1	0	2	1	0	2	1	0
D	2	0	0	2	0	0	2	0	0	2	0	0
E	8	1	0	8	1	0	8	1	0	8	1	0
F	1	0	0	1	0	0	1	0	0	1	0	0
G	1	0	0	1	0	0	1	0	0	1	0	0
H	4	0	0	4	0	0	4	0	0	4	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0
K	4	0	0	4	0	0	4	0	0	4	0	0
L	10	0	0	10	0	0	10	0	0	10	0	0
M	3	1	0	3	1	0	3	1	0	3	1	0
N	0	1	0	0	1	0	0	1	0	0	1	0
P	0	0	0	0	0	0	0	0	0	0	0	0
Q	4	0	0	4	0	0	4	0	0	4	0	0
R	10	1	0	10	1	0	10	1	0	10	1	0
S	4	1	0	4	1	0	4	1	0	4	1	0
T	9	0	0	9	0	0	9	0	0	9	0	0
V	4	0	0	4	0	0	4	0	0	4	0	0
W	5	0	0	5	0	0	5	0	0	5	0	0
Y	3	1	0	3	1	0	3	1	0	3	1	0
Total	h[85]	h[8]	h[0]	h[85]	h[8]	h[0]	h[85]	h[8]	h[0]	h[85]	h[8]	h[0]

Table 2: Amino acid count for 1WBY-H

Amino acid	1WBY:A	1WBY:B	1WBY:C
A	8	1	0
C	1	1	0
D	4	0	0
E	15	0	0
F	1	0	0
G	3	0	0
H	2	0	0
I	1	0	0
K	5	0	0
L	14	0	0
M	5	0	0
N	2	0	0
P	0	0	0
Q	8	0	0
R	8	0	0
S	0	0	0
T	7	1	0
V	2	0	0
W	8	0	0
Y	8	1	0
Total	h[102]	h[4]	h[0]

Table 3: Amino acid count for 2CII-H

Amino acid	2CII:A	2CII:B	2CII:C
A	8	1	0
C	1	1	0
D	4	0	0
E	16	1	0
F	1	0	0
G	3	0	0
H	2	0	0
I	1	0	0
K	5	0	0
L	14	0	0
M	5	0	0
N	2	1	0
P	0	0	0
Q	8	0	0
R	8	1	0
S	0	1	0
T	7	0	0
V	2	0	0
W	9	0	0
Y	8	1	0
Total	h[104]	h[7]	h[0]

Table 4: Amino acid count for 1HSB-H			
Amino acid	1HSB:A	1HSB:B	1HSB:C
A	11	1	0
C	2	1	0
D	2	0	0
E	9	1	0
F	1	0	0
G	2	0	0
H	3	0	0
I	0	0	0
K	4	0	0
L	8	0	0
M	2	0	0
N	1	1	0
P	0	0	0
Q	4	0	0
R	7	1	0
S	3	1	0
T	7	0	0
V	3	0	0
W	6	0	0
Y	3	1	0
Total	h[78]	h[7]	h[0]

Discussion:

The Supplementary Table 1-4 shows the amino acid count for 3MGT, 1WBY, 2CII and 1HSB proteins based on alpha helix secondary structure. For every protein and it's chain the amino acid count will be predicted. The Hydrophobic Amino Acids are Glycine (G), Alanine (A), Valine (V), Leucine (L), Isoleucine (I), Phenylalanine (F), Proline (P) and the Hydrophilic Amino Acids are Asparagine (N), Cysteine (C), Lysine (K), Aspartic (D), Methionine (M), Threonine (T), Glutamine (Q), Histidine (H), Arginine (R), Glutamic (E), Tryptophan (W), Tyrosine (Y), Serine (S). Based on the count of individual amino acid, the wetlab researchers can analyze that which amino acid has majority in that particular protein and also they make decision on that protein whether it belongs to hydrophobic or hydrophilic amino acid. For example, the 1HSB protein consist the count of 25 for Hydrophobic and 53 for Hydophilic. Based on these, the users concluded that the protein belongs to **Hydophilic Acid.** The results have been matched with the PDB results and also it gives the same result. The blue color indicates the amino acid belongs to Hydrophilic and the yellow color belongs to hydrophobic amino acid. From these results, it is inferred that this tool will be most useful for wetlab researchers to identify the properties of the sequence. The PDB results of the 1HSB sequence is shown in Supplementary Fig. 1.

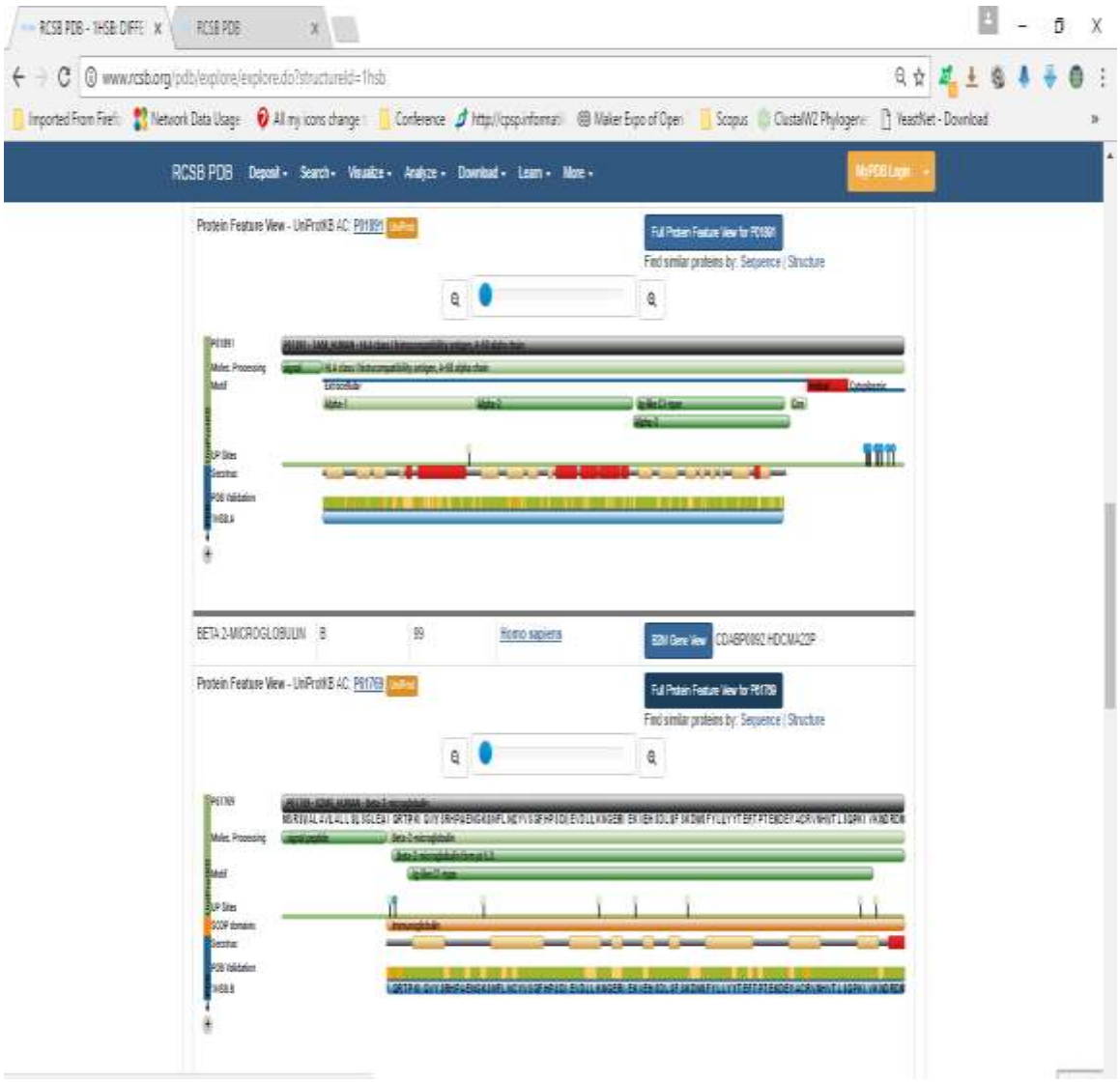

Fig.1: PDB results of the 1HSB sequence

The sample subset of protein details for the three organisms such as Homosapiens, Influenza A Virus Organism and Mus Musculus are shown in Supplementary Table 5-7.

Table 5: Protein Details for the Homosapiens Organism

S.No	Protein Id	Organism	Release Date	Region	Mutation	Experimental Method
1	1aqd	Homo Sapiens	1998-09-16	-	-	X-ray
2	1dlh	Homo Sapiens	1994-06-22	England	-	X-Ray
3	1fyt	Homo Sapiens	2000-11-08	-	yes	X-Ray
4	1h15	Homo Sapiens	2002-10-03	-	-	X-Ray
5	1hhi	Homo Sapiens	1993-10-31	-	-	X-Ray
6	1hhk	Homo Sapiens	1993-10-31	-	-	X-Ray

7	1hsb	Homo Sapiens	1993-10-31	-	-	X-Ray
8	1hxy	Homo Sapiens	2001-06-27	Equine-USA	-	X-Ray
9	1kg0	Homo Sapiens	2002-03-27	-	-	X-Ray
10	1lo5	Homo Sapiens	2002-12-18	-	yes	X-Ray
11	1pyw	Homo Sapiens	2003-12-09		yes	X-Ray
12	1r5i	Homo Sapiens	2004-03-16	Swine-Hong Kong	-	X-Ray
13	2bst	Homo Sapiens	2005-05-24	-	-	X-Ray
14	2cii	Homo Sapiens	2006-03-29	-	-	X-Ray
15	2f0z	Homo Sapiens	2006-11-21	-	-	X-Ray
16	2f10	Homo Sapiens	2006-11-21	-	-	X-Ray
17	2f11	Homo Sapiens	2006-11-21	-	-	X-Ray
18	2f12	Homo Sapiens	2006-11-21	-	-	X-Ray
19	2f13	Homo Sapiens	2006-11-21	-	-	X-Ray
20	2g9h	Homo Sapiens	2006-07-11	England	-	X-Ray
21	2icw	Homo Sapiens	2007-03-20	-	yes	X-Ray
22	2ipk	Homo Sapiens	2007-03-13	Mallard-Sweden	yes	X-Ray
23	2jdq	Homo Sapiens	2007-02-27	-	-	X-Ray
24	2oje	Homo Sapiens	2007-01-23	Taiwan	-	X-Ray
25	2rhk	Homo Sapiens	2008-07-01	-	yes	X-Ray
26	2vlj	Homo Sapiens	2008-01-22	-	-	X-Ray
27	2vlk	Homo Sapiens	2008-01-22	-	-	X-Ray
28	2vll	Homo Sapiens	2008-01-22	-	-	X-Ray
29	2vlr	Homo Sapiens	2008-01-22	-	-	X-Ray
30	2wbj	Homo Sapiens	2009-04-07	-	-	X-Ray
31	2x4n	Homo Sapiens	2010-03-02	-	-	X-Ray
32	2x4p	Homo Sapiens	2010-03-02	-	-	X-Ray
33	2x4q	Homo Sapiens	2010-03-02	-	-	X-Ray
34	2x4s	Homo Sapiens	2010-03-02	-	-	X-Ray
35	2x70	Homo Sapiens	2010-12-08	-	-	X-Ray
36	2xn9	Homo Sapiens	2010-11-24	-	yes	X-Ray
37	3fkq	Homo Sapiens	2009-01-06	-	-	X-Ray
38	3g81	Homo Sapiens	2009-03-17	-	-	X-Ray
39	3g83	Homo Sapiens	2009-03-17	-	-	X-Ray
40	3g84	Homo Sapiens	2009-03-17	-	yes	X-Ray
41	3gbm	Homo Sapiens	2009-03-10	Vietnam	-	X-Ray
42	3gbn	Homo Sapiens	2009-03-10	Brevig mission	-	X-Ray

43	3lkn	Homo Sapiens	2010-07-07	-	-	X-Ray
44	3lko	Homo Sapiens	2010-07-07	-	-	X-Ray
45	3lkp	Homo Sapiens	2010-07-07	-	-	X-Ray
46	3lkq	Homo Sapiens	2010-07-07	-	-	X-Ray
47	3lkr	Homo Sapiens	2010-07-07	-	-	X-Ray
48	3lks	Homo Sapiens	2010-07-07	-	-	X-Ray
49	3lzf	Homo Sapiens	2010-04-07	South Carolina	-	X-Ray
50	3mgo	Homo Sapiens	2010-05-19	-	-	X-Ray

Table 6: Protein Details for the Influenza A Virus Organism

S.No	Protein id	Organism	Release date	Region	Mutation	Experimental method
1	2wfs	Influenza A virus	2009-07-07	-	-	Electron Microscopy
2	1ns1	Influenza A virus	1998-01-14	-	-	Solution NMR
3	2jrd	Influenza A virus	2007-07-10	-	Yes	Solution NMR
4	2l24	Influenza A virus	2011-06-22	-	Yes	Solution NMR
5	1a14	Influenza A virus	1998-05-13	-	-	X-Ray
6	1bji	Influenza A virus	1998-06-03	-	-	X-Ray
7	1ea3	Influenza A virus	2001-04-26	-	-	X-Ray
8	1ha0	Influenza A virus	1998-10-12	-	Yes	X-Ray
9	1hhi	Influenza A virus	1993-10-31	-	-	X-Ray
10	1ing	Influenza A virus	1996-08-17	-	Yes	X-Ray
11	1inh	Influenza A virus	1996-08-17	-	Yes	X-Ray
12	1inw	Influenza A virus	1995-02-07	-	Yes	X-Ray
13	1inx	Influenza A virus	1995-02-07	-	Yes	X-Ray
14	1iny	Influenza A virus	1995-02-07	-	Yes	X-Ray
15	1jsm	Influenza A virus	2001-09-26	-	-	X-Ray
16	1jsn	Influenza A virus	2001-09-26	-	-	X-Ray
17	1jso	Influenza A virus	2001-09-26	-	-	X-Ray
18	1l7f	Influenza A virus	2002-05-29	-	-	X-Ray
19	1l7g	Influenza A virus	2002-05-29	-	Yes	X-Ray
20	1l7h	Influenza A virus	2002-05-29	-	Yes	X-Ray
21	1mql	Influenza A virus	2003-08-26	-	-	X-Ray
22	1mqm	Influenza A virus	2003-08-26	-	-	X-Ray
23	1mqn	Influenza A virus	2003-08-26	-	-	X-Ray
24	1mwe	Influenza A virus	1998-03-04	-	-	X-Ray
25	1nma	Influenza A virus	1995-09-15	-	Yes	X-Ray
26	1nmb	Influenza A virus	1995-09-15	-	Yes	X-Ray
27	1nmc	Influenza A virus	1998-09-23	-	-	X-Ray
28	1nna	Influenza A virus	1994-04-30	-	-	X-Ray
29	1nnb	Influenza A virus	1994-04-30	-	-	X-Ray
30	1pd3	Influenza A virus	2003-12-16	-	-	X-Ray

31	1qu1	Influenza A virus	2000-01-05	-	Yes	X-Ray
32	1rd8	Influenza A virus	2004-03-23	-	-	X-Ray
33	1rv0	Influenza A virus	2004-03-30	-	-	X-Ray
34	1ti8	Influenza A virus	2005-06-21	-	Yes	X-Ray
35	1v0z	Influenza A virus	2006-01-25	-	-	X-Ray
36	1w1x	Influenza A virus	2006-01-25	-	-	X-Ray
37	1w20	Influenza A virus	2006-01-25	-	-	X-Ray
38	1w21	Influenza A virus	2006-01-25	-	-	X-Ray
39	1wbx	Influenza A virus	2005-01-19	-	-	X-Ray
40	1wby	Influenza A virus	2005-01-19	-	-	X-Ray
41	1wbz	Influenza A virus	2005-01-19	-	-	X-Ray
42	1xoe	Influenza A virus	2005-01-11	-	-	X-Ray
43	1xog	Influenza A virus	2005-01-11	-	-	X-Ray
44	2aeq	Influenza A virus	2005-12-20	-	-	X-Ray
45	2b8h	Influenza A virus	2006-09-05	-	Yes	X-Ray
46	2bst	Influenza A virus	2005-05-24	-	-	X-Ray
47	2c4a	Influenza A virus	2007-03-27	-	-	X-Ray
48	2c4l	Influenza A virus	2007-03-27	-	-	X-Ray
49	2cml	Influenza A virus	2007-06-05	-	-	X-Ray
50	2ht5	Influenza A virus	2006-09-05	-	-	X-Ray

Table 7: Protein Details for the Mus Musculus Organism

S.No	Protein Id	Organism	Release Date	Region	Mutation	Experimental Method
1	1a14	Mus Musculus	1998-05-13	-	-	X-Ray
2	1eo8	Mus Musculus	2000-04-12	-	-	X-Ray
3	1frg	Mus Musculus	1994-05-31	Assam	-	X-Ray
4	1hil	Mus Musculus	1994-01-31	-	-	X-Ray
5	1him	Mus Musculus	1994-01-31	-	-	X-Ray
6	1hin	Mus Musculus	1994-01-31	-	-	X-Ray
7	1hoc	Mus Musculus	1994-04-30	Hong Kong	-	X-Ray
8	1ifh	Mus Musculus	1993-10-31	-	-	X-Ray
9	1ken	Mus Musculus	2002-04-24	-	-	X-Ray
10	1nca	Mus Musculus	1994-01-31	Australia	-	X-Ray
11	1ncb	Mus Musculus	1994-01-31	Australia	-	X-Ray
12	1ncc	Mus Musculus	1994-01-31	Australia	-	X-Ray
13	1ncd	Mus Musculus	1994-01-31	Maine	-	X-Ray
14	1nma	Mus Musculus	1995-09-15	-	-	X-Ray
15	1nmb	Mus Musculus	1995-09-15	-	-	X-Ray
16	1nmc	Mus Musculus	1998-09-23	-	-	X-Ray
17	1qfu	Mus Musculus	1999-04-16	-	-	X-Ray
18	1wbx	Mus Musculus	2005-01-19	-	-	X-Ray
19	1wby	Mus Musculus	2005-01-19	-	-	X-Ray

20	1wbz	Mus Musculus	2005-01-19	-	-	X-Ray
21	1yn6	Mus Musculus	2005-06-28	Swine-lowa	-	X-Ray
22	1yn7	Mus Musculus	2005-06-28	Swine-Lowa	Yes	X-Ray
23	1zt1	Mus Musculus	2005-10-18	-	-	X-Ray
24	2aep	Mus Musculus	1994-01-31	-	-	X-Ray
25	2aeq	Mus Musculus	1997-09-04	-	-	X-Ray
26	2cii	Mus Musculus	2006-03-29	-	-	X-Ray
27	2clv	Mus Musculus	2006-06-14	-	Yes	X-Ray
28	2clz	Mus Musculus	2006-06-14	-	Yes	X-Ray
29	2fwo	Mus Musculus	2006-02-21	England	-	X-Ray
30	2icw	Mus Musculus	2007-03-20	-	Yes	X-Ray
31	2ve6	Mus Musculus	2008-01-22	-	-	X-Ray
32	2vir	Mus Musculus	1998-04-29	-	-	X-Ray
33	2vis	Mus Musculus	1998-04-29	-	Yes	X-Ray
34	2vit	Mus Musculus	1998-04-29	-	Yes	X-Ray
35	3buy	Mus Musculus	2008-03-25	-	-	X-Ray
36	3cpl	Mus Musculus	2008-11-18	-	-	X-Ray
37	3ftg	Mus Musculus	2009-12-29	-	-	X-Ray
38	3pqy	Mus Musculus	2011-05-18	Swine-Lowa	-	X-Ray
39	4apq	Mus Musculus	2013-04-24	-	-	X-Ray
40	4f15	Mus Musculus	2013-05-15	Korea	-	X-Ray
41	4gms	Mus Musculus	2012-10-03	Victoria	-	X-Ray
42	4gmt	Mus Musculus	2012-10-03	-	-	X-Ray
43	4hlz	Mus Musculus	2013-04-17	Japan	-	X-Ray
44	4huu	Mus Musculus	2013-02-27	Swine-Korea	-	X-Ray
45	4huv	Mus Musculus	2013-02-27	Swine-Korea-	-	X-Ray
46	4huw	Mus Musculus	2013-02-27	Swine-Korea	-	X-Ray
47	4hux	Mus Musculus	2013-02-27	Swine-Korea	-	X-Ray
48	4hv8	Mus Musculus	2013-02-27	Swine-Korea	-	X-Ray
49	4l8b	Mus Musculus	2013-10-16	-	-	X-Ray
50	4l8c	Mus Musculus	2013-10-16	-	-	X-Ray