Predicting the host of influenza viruses based on the word vector

Beibei Xu [1], **Zhiying Tan** [1], **Kenli Li** [1], **Taijiao Jiang** Corresp., [2], **Yousong Peng** Corresp. [3]

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[2] Center of System Medicine, Institute of Basic Medical Sciences, Beijing, China
[3] College of Biology, Hunan University, Changsha, China

Corresponding Authors: Taijiao Jiang, Yousong Peng
Email address: taijiao@ibms.pumc.edu.cn, pys2013@hnu.edu.cn

Newly emerging influenza viruses continue to threaten public health. A rapid determination of the host range of newly discovered influenza viruses would assist in early assessment of their risk. Here, we attempted to predict the host of influenza viruses using the Support Vector Machine (SVM) classifier based on the word vector, a new representation and feature extraction method for biological sequences. The results show that the length of word within the word vector, the sequence type (DNA or protein) and the species from which the sequences were derived for generating the word vector all influence the performance of models in predicting the host of influenza viruses. In nearly all cases, the models built on the surface proteins hemagglutinin (HA) and neuraminidase (NA) (or their genes) produced better results than internal influenza proteins (or their genes). The best performance was achieved when the model was built on the HA gene based on word vectors (words of three-letters long) generated from DNA sequences of the influenza virus. This results in accuracies of 99.7% for avian, 96.9% for human and 90.6% for swine influenza viruses. Compared to the method of sequence homology best-hit searches using Basic Local Alignment Search Tool (BLAST), the word vector-based models still need further improvements in predicting the host of influenza A viruses.

Predicting the host of influenza viruses based on the word vector

Beibei Xu[1], Zhiying Tan[1], Kenli Li[1], Taijiao Jiang[2, 3, *], Yousong Peng[4,*]

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha, 410082, China

[2] Center of System Medicine, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, 100005, China

[3] Suzhou Institute of Systems Medicine, Suzhou, Jiangsu, 215123, China

[4] College of Biology, Hunan University, Changsha, 410082, China

* To whom correspondence should be addressed. Email: taijiao@moon.ibp.ac.cn(TJ) or pys2013@hnu.edu.cn(YP)

Abstract

Newly emerging influenza viruses continue to threaten public health. A rapid determination of the host range of newly discovered influenza viruses would assist in early assessment of their risk. Here, we attempted to predict the host of influenza viruses using the Support Vector Machine (SVM) classifier based on the word vector, a new representation and feature extraction method for biological

sequences. The results show that the length of word within the word vector, the sequence type (DNA or protein) and the species from which the sequences were derived for generating the word vector all influence the performance of models in predicting the host of influenza viruses. In nearly all cases, the models built on the surface proteins hemagglutinin (HA) and neuraminidase (NA) (or their genes) produced better results than internal influenza proteins (or their genes). The best performance was achieved when the model was built on the HA gene based on word vectors (words of three-letters long) generated from DNA sequences of the influenza virus. This results in accuracies of 99.7% for avian, 96.9% for human and 90.6% for swine influenza viruses. Compared to the method of sequence homology best-hit searches using Basic Local Alignment Search Tool (BLAST), the word vector-based models still need further improvements in predicting the host of influenza A viruses.

Introduction

The influenza virus is a negative-sense, single-stranded, segmented RNA virus. Its genome is composed of eight segments and mainly encodes twelve proteins, including two surface proteins HA and NA, and ten internal proteins PB2, PB1, PA, NP, M1, M2, NS1, NS2, PA-X and PB1-F2. Influenza viruses could be mainly

38 separated into types A, B and C, while type A could be further separated into
39 subtypes according to the HA and NA proteins, such as H3N2, H1N1, H5N1, and
40 so on (Taubenberger & Kash 2010). Type B and C influenza viruses mainly infect
41 humans, whereas type A can infect a wide range of species, such as birds (poultry)
42 and mammals (pigs, bats) including humans (Webster et al. 1992). Among them,
43 avian, human and swine influenza viruses are most commonly observed, and cause
44 large health and economic loss to human society. In recent years, human infections
45 by what were considered typical avian and swine strains have become more
46 common, for instance infections due to influenza H7N9 and H5N8 (Peiris et al.
47 2016; Su et al. 2015). Rapid determination of the host range of a given influenza
48 virus could assist in early evaluation of the potential risk of emerging subtypes.

49 Avian influenza virus is considered to be the evolutionary ancestor of all other
50 influenza viruses (Webster et al. 1992). When an avian influenza virus infects a
51 different host species and is able to spread within this new host, mutations rapidly
52 accumulate as the viral population adapts to this host. This situation has been
53 observed in human and swine influenza viruses. Several molecular markers for
54 human or other influenza viruses have been identified at the amino acid level,
55 either experimentally or by means of computational analyses (Chen et al. 2006;
56 Finkelstein et al. 2007; Kim et al. 2010; Tamuri et al. 2009). For example, a Lysine
57 in position 627 of internal protein PB2 tends to be favored in human strains, while

avian strains usually have a Glutamic acid at this position (Finkelstein et al. 2007; Kim et al. 2010). Several studies further attempted to classify the host of influenza virus by machine learning methods (ElHefnawi & Sherif 2014; Sherif et al. 2011; Attaluri et al. 2010). For example, Attaluri et al. integrated multiple machine learning techniques to predict the host of influenza A viruses and achieved accuracies ranging from 0.84 to 0.98. However, most of these studies either used a selection of HA subtypes only, or a relatively small dataset, ignoring the real, extensive genetic diversity of influenza viruses.

In this work, we used the largest influenza dataset known to date, including 163666 unique DNA and 150947 unique protein sequences, to predict the host of influenza viruses based on the nucleotide and amino acid word vector. The word vector is a new representation and feature extraction method (Mikolov et al. 2013), which was originally developed and used in natural language processing. It was first applied to biological research in 2015 (Asgari & Mofrad 2015) and proved to be useful for protein family classification and disordered protein prediction. Here, we applied word vectors to predict the host of influenza viruses and achieved an overall accuracy of 0.97, which further demonstrated its strength in biological sequence representation.

Materials and Methods

Overview of this work

Figure 1 shows the workflow of this work. Firstly, all protein and DNA sequences of influenza viruses (denoted as influenza protein and DNA dataset respectively) and all non-redundant protein sequences of all species (denoted as SwissProt dataset) were collected, which were used to generate word vectors by the tool word2vec. Then, all non-redundant protein and DNA sequences of influenza A viruses with known host (avian, human and swine) were transformed into protein and DNA vectors based on word vectors. Finally, the Support Vector Machine (SVM) models were built for classifying influenza viruses of avian/human, avian/swine and human/swine based on protein or DNA vectors. A voting strategy was used to determine the predicted host for the influenza virus.

Datasets

The SwissProt dataset were derived from the Swiss-Prot database on UniProt (UniProt 2016) on April 11th, 2016. It contains 550740 protein sequences with a length ranging from 11 to 35213 amino acids (aa).

For the influenza DNA dataset, the nucleotide (nt) sequences for eight genes, including HA and NA as well as internal protein genes PB2, PB1, PA, NP, MP and

NS, of influenza A viruses with known host (avian, human and swine) were extracted from the database of Influenza Virus Resources (Bao et al. 2008) on April 26th, 2016. At the same time, amino acid sequences of the 12 proteins (HA and NA, and the ten internal proteins) were extracted to produce an influenza protein dataset. In total, 385788 DNA and 607327 protein sequences were collected. To reduce the computational cost, redundancy was removed at 100% level with the help of the software package cd-hit (Li & Godzik 2006). This step maintained 163666 unique DNA and 150947 unique protein sequences for analysis. The number of protein sequences used in this study for each protein of avian, human and swine influenza viruses is shown in Table 1; the number of genes is included in Table S1. These datasets are much larger than those used in previous studies.

Word vector generation and protein sequence vectorization

The tool word2vec is a software package developed for producing word embeddings (Mikolov et al. 2013). It takes a large corpus of text (here, it refers to large number of protein or DNA sequences which were separated into words) as its input and outputs a vector space of several hundred dimensions. Each unique word in the texts is assigned a corresponding vector in the space, during which they are closely located to other words that share a common context. Here, the tool word2vec was used to generate the word vectors of 200 dimensions using the

SwissProt dataset and the influenza datasets respectively. The skip-gram model and hierarchical softmax algorithm were used in the word2vec, with other parameters in default values. The word vectors with words of two to four amino acids (or nucleotides) long were all generated in the same way.

The vectorization of protein (or DNA) sequences was adapted from Asgari and Mofrad's work (Asgari & Mofrad 2015). Firstly, each protein (or DNA) sequence was separated into overlapping words of N (2~4) amino acids (or nucleotides). Then, the word vectors for all these words were summed up and averaged, which led to the protein (or DNA) vectors of 200-dimensions for the protein (or DNA) sequences.

Predicting the host for the influenza virus with SVM

The SVM models for predicting the host of influenza viruses were built using functions of svmtrain() and svmclassify() (Change & Lin 2011) in MATLAB R2014b. The Gaussian Radial Basis Function kernel "rbf" with default parameters was used for the SVM models. Three SVM models were built to discriminate the influenza viruses of avian and human, avian and swine, human and swine based on word vectors. A simple voting strategy was used to determine the final prediction for the host of influenza viruses. Ten-fold cross-validations were used to evaluate the performance of the SVM models.

Predicting the host of influenza viruses with sequence homology search

The method of Profile hidden Markov model (HMM) through the package of HMMER3 (Eddy 2010), and the method of Basic Local Alignment Search Tool (BLAST) through the package of BLAST+ (Altschul et al. 1990), were used for inferring the host of influenza A viruses based on homologies of protein or DNA sequences. For each gene or protein, 75% of protein (or DNA) sequences were randomly selected for building the library (for BLAST) or profile (for HMM), while the remaining protein (or DNA) sequences were used to test through the best hit search.

Results

Predict the host of influenza viruses based on word vectors derived from influenza protein dataset

We firstly attempted to predict the host of influenza A viruses based on word vectors derived from influenza protein dataset. Figure 2 shows that the SVM models built on the surface proteins HA and NA, and on the internal proteins PB2, PB1, PA and NP, performed much better than those built on other internal proteins did (including M1, M2, NS1, NS2, PB1-F2 and PA-X). The overall accuracies ranged from 0.79~0.96 (summarized in Table S2). The length of word in the word vector has a significant influence on the model's performances: the models based

153 on four-letters words performed best for all proteins. Further analyses on the
154 model's performance by host show that all models predict most accurately for the
155 avian influenza virus, with accuracies ranging from 0.97~1. For human and swine
156 influenza viruses, the models achieved accuracies of approximately 0.9 for HA,
157 NA, PB2, PB1, PA and NP, while for the other proteins performance was rather
158 poor.

159 **Predict the host of influenza viruses based on word vectors derived from the**
160 **SwissProt dataset**

161 We next investigated the influence of species of protein sequence, which were used
162 to generate the word vector, on the prediction of the host of influenza viruses. The
163 SwissProt dataset included protein sequences of virus, bacteria, fungi, plant,
164 animal, and so on. In theory, the word vectors derived from the influenza protein
165 dataset should reflect more accurately the influenza virus than those derived from
166 the SwissProt dataset. Figure 3 shows the overall accuracies for the SVM models
167 based on two kinds of word vectors with words of two to four amino acids long.
168 The models based on two kinds of word vectors achieved comparable
169 performances. For the word vector with words of two to three letters long, the
170 models based on word vectors derived from the SwissProt dataset even
171 outperformed those based on word vectors derived from the influenza protein
172 dataset. However, the best performance (overall accuracy greater than 0.96) was

achieved on the model built on the HA protein, which used word vectors with words of four-letters long derived from the influenza protein dataset (Table S2 and S3). Besides, accuracies of models based on word vectors derived from the SwissProt dataset decreased as the length of the word in the word vector, while the opposite trend was observed for models based on those derived from the influenza protein dataset (Figure 3).

Predict the host of influenza viruses based on word vectors derived from influenza DNA dataset

Then, we continued to investigate the influence of data type, DNA or protein, on predicting the host of influenza viruses based on the word vector. The influenza DNA dataset were used to generate word vectors with words of two to four nucleotides long. Figure 4 shows that in most cases, models based on word vectors derived from DNA sequences outperformed those based on word vectors derived from protein sequences. Excellent performance was obtained with a DNA word length of three, with overall accuracies greater than 0.95 for most genes (Table S4). As before, the best performance was achieved on the model based on word vectors of HA gene, with an overall prediction accuracy of 0.97. More specifically, the prediction accuracy for avian, human and swine influenza viruses equaled 0.997, 0.969 and 0.906, respectively (Table S4).

192 **Predict the host of influenza viruses based on sequence homology search**

193 Sequence homology search through the methods of BLAST and HMM can also be
194 used for inferring the host of viruses. Here, we tested both methods in inferring the
195 host of influenza A viruses based on both protein and DNA sequences. As shown
196 in Table 2, when using the protein sequence, the method based on word vectors
197 outperformed those based on HMM and BLAST, at least for HA, NA, and the
198 proteins PB2, PB1, PA, NP and M2. For the other internal proteins (including the
199 short proteins) the reverse was the case. For the methods based on sequence
200 homology search, BLAST performed slightly better than HMM did, especially for
201 HA and NA. Surprisingly, when using the DNA sequence, the method based on
202 BLAST outperformed all the other methods for nearly all genes (Table S5). It
203 achieved an overall accuracy of 0.979 on the HA gene, which is greater than that of
204 all the other models tested here.

205 **Discussion**

206 This work investigated the prediction of the host of influenza A viruses based on
207 word vectors. For all genes or proteins, the predictions for avian influenza viruses
208 were more accurate than for human or swine influenza viruses. This may partly be
209 caused by occasional infections of humans or swine by what actually were avian
210 influenza viruses (Beigel et al. 2005; Claas et al. 1998), which may have weakened

any host-specific signal in the non-avian hosts.

The surface proteins HA and NA were observed to be better discriminators for the host of influenza viruses than the internal viral proteins. This is most likely related to the selective pressure posed by the host: since HA and NA are the main antigens of influenza virus they will be recognized by the immune system of the host (Couch & Kasel 1983). Therefore, these proteins have to mutate rapidly to maintain a stable population in a new host, a mechanism of host adaptation that will lead to divergence of lineages. Surprisingly, the results obtained with internal proteins PB2, PB1, PA and NP were comparable to those of the surface proteins. These proteins constitute the RNA polymerase complex of the virus, which is responsible for RNA replication and is thus directly responsible for the survivor ability of the virus (te Velthuis & Fodor 2016). All viral proteins are translated by the host, but these proteins are most important for rapid viral reproduction, thus they will also adapt rapidly to the new host (Mehle & Doudna 2009).

As it is known, the HA proteins diverge most among all the proteins of influenza A viruses, which results in 18 HA subtypes reported until now (Tong et al. 2013). The best performance achieved on the model of HA protein suggests that the word vector may capture the intrinsic difference between the influenza virus of different hosts, irrespective of the HA subtypes.

The word vectors could be generated with protein or DNA sequences of any species. In theory, the word vector derived from the influenza protein dataset could reflect more accuracy the influenza virus than those derived from the SwissProt dataset. However, in most cases, the models based on the former performed comparably with those based on the latter (Figure 3). This may reflect the similar principles in organizing amino acids into protein sequences in all species.

A limitation of this work is that only the word vector was used in predicting the host of influenza viruses. More features such as the amino acids composition, motif frequency and molecular markers may be integrated to improve the accuracy of models in predicting the human and swine influenza viruses. Overall, this work should be an interesting attempt in using the word vector in biological sequence representation. The models based on word vectors achieved high accuracies in predicting the host of influenza viruses, which may be helpful in influenza prevention.

References

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403-410.

Asgari E, and Mofrad MR. 2015. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLoS One* 10:e0141287. 10.1371/journal.pone.0141287

250 Attaluri PK, Chen ZX, Lu GQ. 2010. Applying neural networks to classify influenza virus antigenic types and hosts.
251 *2010 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*
252 (CIBCB): IEEE.
253 Bao YM, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, and Lipman D. 2008. The
254 influenza virus resource at the national center for biotechnology information. *J Virol* 82:596-601. Doi
255 10.1128/Jvi.02005-07
256 Beigel JH, Farrar J, Han AM, Hayden FG, Hyer R, de Jong MD, Lochindarat S, Nguyen TK, Nguyen TH, Tran TH,
257 Nicoll A, Touch S, and Yuen KY. 2005. Avian influenza A (H5N1) infection in humans. *N Engl J Med*
258 353:1374-1385. 10.1056/NEJMra052211
259 Chang CC and Lin CJ. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent*
260 *Systems and Technology*, 2:27:1--27:27.
261 Chen GW, Chang SC, Mok CK, Lo YL, Kung YN, Huang JH, Shih YH, Wang JY, Chiang C, Chen CJ, and Shih SR.
262 2006. Genomic signatures of human versus avian influenza A viruses. *Emerging Infectious Diseases*
263 12:1353-1360. 10.3201/eid1209.060276
264 Claas EC, Osterhaus AD, van Beek R, De Jong JC, Rimmelzwaan GF, Senne DA, Krauss S, Shortridge KF, and
265 Webster RG. 1998. Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus.
266 *Lancet* 351:472-477. 10.1016/S0140-6736(97)11212-0
267 Couch RB, and Kasel JA. 1983. Immunity to influenza in man. *Annu Rev Microbiol* 37:529-549.
268 10.1146/annurev.mi.37.100183.002525
269 ElHefnawi M, and Sherif FF. 2014. Accurate classification and hemagglutinin amino acid signatures for influenza A
270 virus host-origin association and subtyping. *Virology* 449:328-338. 10.1016/j.virol.2013.11.010
271 Eddy S. 2010. HMMER3: a new generation of sequence homology search software. *Available at http://hmmer.org/*.
272 Finkelstein DB, Mukatira S, Mehta PK, Obenauer JC, Su X, Webster RG, and Naeve CW. 2007. Persistent host
273 markers in pandemic and H5N1 influenza viruses. *J Virol* 81:10292-10299. 10.1128/JVI.00921-07
274 Kim JH, Hatta M, Watanabe S, Neumann G, Watanabe T, and Kawaoka Y. 2010. Role of host-specific amino acids
275 in the pathogenicity of avian H5N1 influenza viruses in mice. *J Gen Virol* 91:1284-1289.
276 10.1099/vir.0.018143-0
277 Li W, and Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide

sequences. *Bioinformatics* 22:1658-1659. 10.1093/bioinformatics/btl158

Mehle A, and Doudna JA. 2009. Adaptive strategies of the influenza virus polymerase for replication in humans. *Proc Natl Acad Sci U S A* 106:21312-21316. 10.1073/pnas.0911915106

Mikolov T, Chen K, Corrado GS, Dean J. 2013. Efficient Estimation of Word Representations in Vector Space. *In Proceedings of Workshop at ICLR*, 2013.

Peiris JS, Cowling BJ, Wu JT, Feng L, Guan Y, Yu H, and Leung GM. 2016. Interventions to reduce zoonotic and pandemic risks from avian influenza in Asia. *Lancet Infectious Diseases* 16:252-258. 10.1016/S1473-3099(15)00502-2

Sherif FF, Kadah Y, El-Hefnawi M. 2011. Classification of human vs. non-human, and subtyping of human influenza viral strains using Profile Hidden Markov Models. 2011. *1st Middle East Conference on Biomedical Engineering* (MECBME): IEEE.

Su S, Bi Y, Wong G, Gray GC, Gao GF, and Li S. 2015. Epidemiology, Evolution, and Recent Outbreaks of Avian Influenza Virus in China. *J Virol* 89:8671-8676. 10.1128/JVI.01034-15

Tamuri AU, Dos Reis M, Hay AJ, and Goldstein RA. 2009. Identifying changes in selective constraints: host shifts in influenza. *PLoS Comput Biol* 5:e1000564. 10.1371/journal.pcbi.1000564

Taubenberger JK, and Kash JC. 2010. Influenza virus evolution, host adaptation, and pandemic formation. *Cell Host Microbe* 7:440-451. 10.1016/j.chom.2010.05.009

te Velthuis AJW, and Fodor E. 2016. Influenza virus RNA polymerase: insights into the mechanisms of viral RNA synthesis. *Nature Reviews Microbiology* 14:479-493. 10.1038/nrmicro.2016.87

Tong SX, Zhu XY, Li Y, Shi M, Zhang J, Bourgeois M, Yang H, Chen XF, Recuenco S, Gomez J, Chen LM, Johnson A, Tao Y, Dreyfus C, Yu WL, McBride R, Carney PJ, Gilbert AT, Chang J, Guo Z, Davis CT, Paulson JC, Stevens J, Rupprecht CE, Holmes EC, Wilson IA, and Donis RO. 2013. New World Bats Harbor Diverse Influenza A Viruses. *PLoS Pathog* 9. ARTN e100365710.1371/journal.ppat.1003657

UniProt. 2016. *Available at http://www.uniprot.org/*.

Webster RG, Bean WJ, Gorman OT, Chambers TM, and Kawaoka Y. 1992. Evolution and ecology of influenza A viruses. *Microbiol Rev* 56:152-179.

305

306 **Figure Legends**

307 **Table 1** The number of non-redundant sequences used in this study for each
308 protein of avian, human and swine influenza A viruses.

Protein	Avian	Human	Swine	Protein	Avian	Human	Swine
HA	15328	17872	7132	M1	1717	1134	1011
NA	10295	9834	4802	M2	2360	1456	1431
PB2	8134	4363	2386	NS1	5364	3128	2076
PB1	7323	4026	2343	NS2	2231	1000	995
PA	7925	4173	2461	PB1-F2	4151	1180	934
NP	4816	2261	1929	PA-X	1716	738	922

309

310 **Table 2** Comparison of methods for predicting the host of influenza A viruses
311 based on the word vector and based on sequence homology best-hit searches using
312 protein sequences. The table listed the overall accuracies for predicting the host of
313 influenza A viruses. For the method based on word vectors, the optimal model for
314 individual proteins was used.

315

Protein	Methods for predicting the host of influenza A virus		
	Word Vector	BLAST	HMM
HA	0.964	0.950	0.676
NA	0.955	0.914	0.593
PB2	0.931	0.892	0.885
PB1	0.928	0.898	0.798
PA	0.933	0.917	0.822
NP	0.912	0.837	0.830
M1	0.712	0.676	0.672
M2	0.509	0.807	0.867
NS1	0.799	0.864	0.895
NS2	0.561	0.748	0.866
PB1-F2	0.712	0.952	0.955
PA-X	0.625	0.896	0.730

316 **Figure 1** The workflow of the methodological approach used. For explanation see

317 text.

318

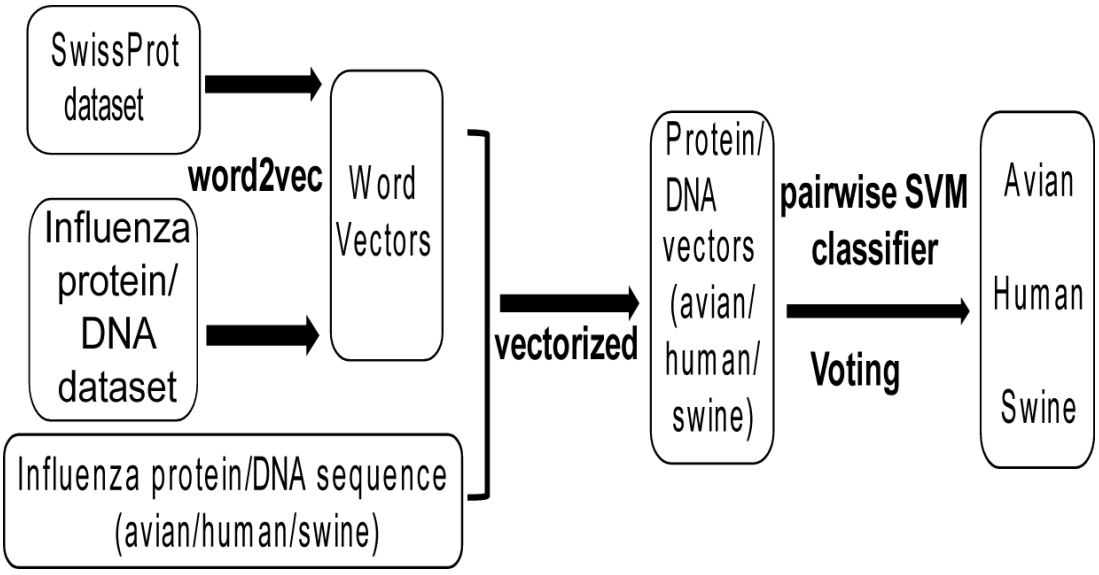
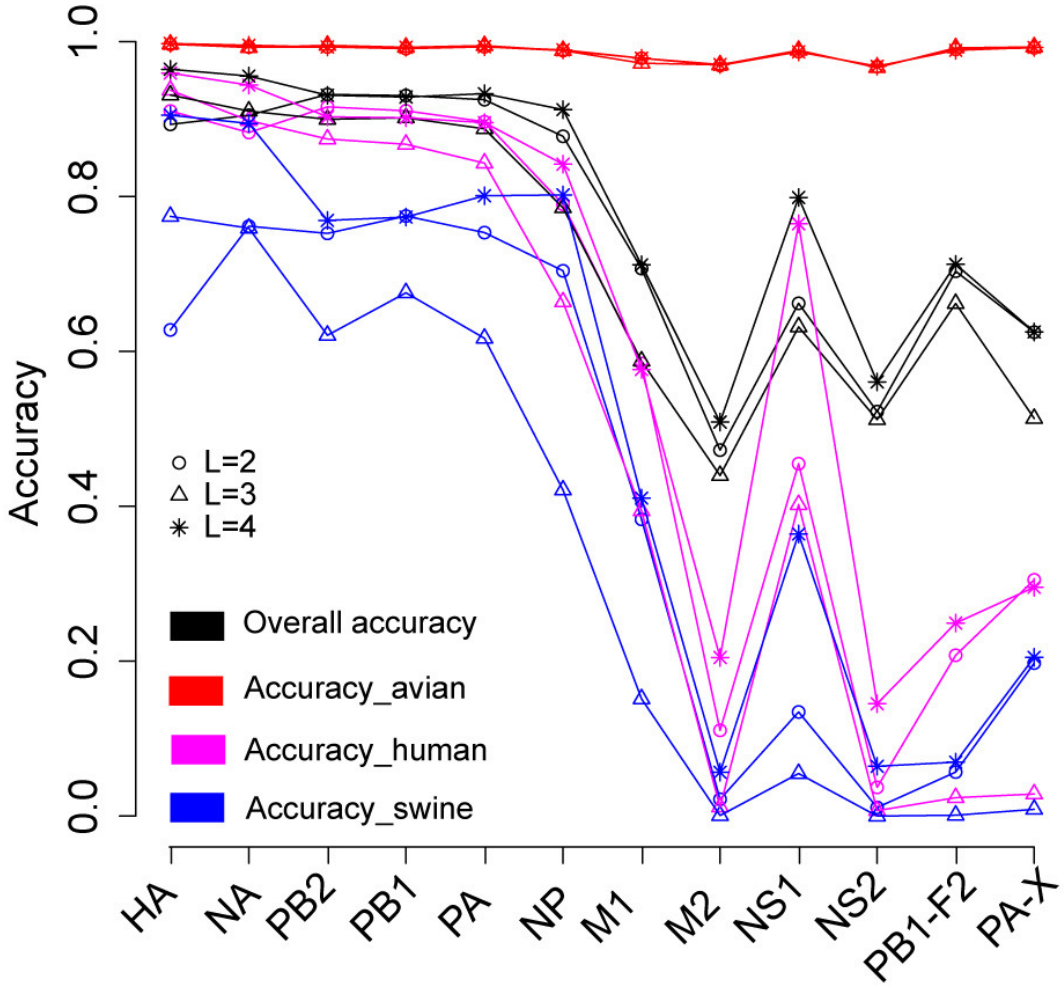
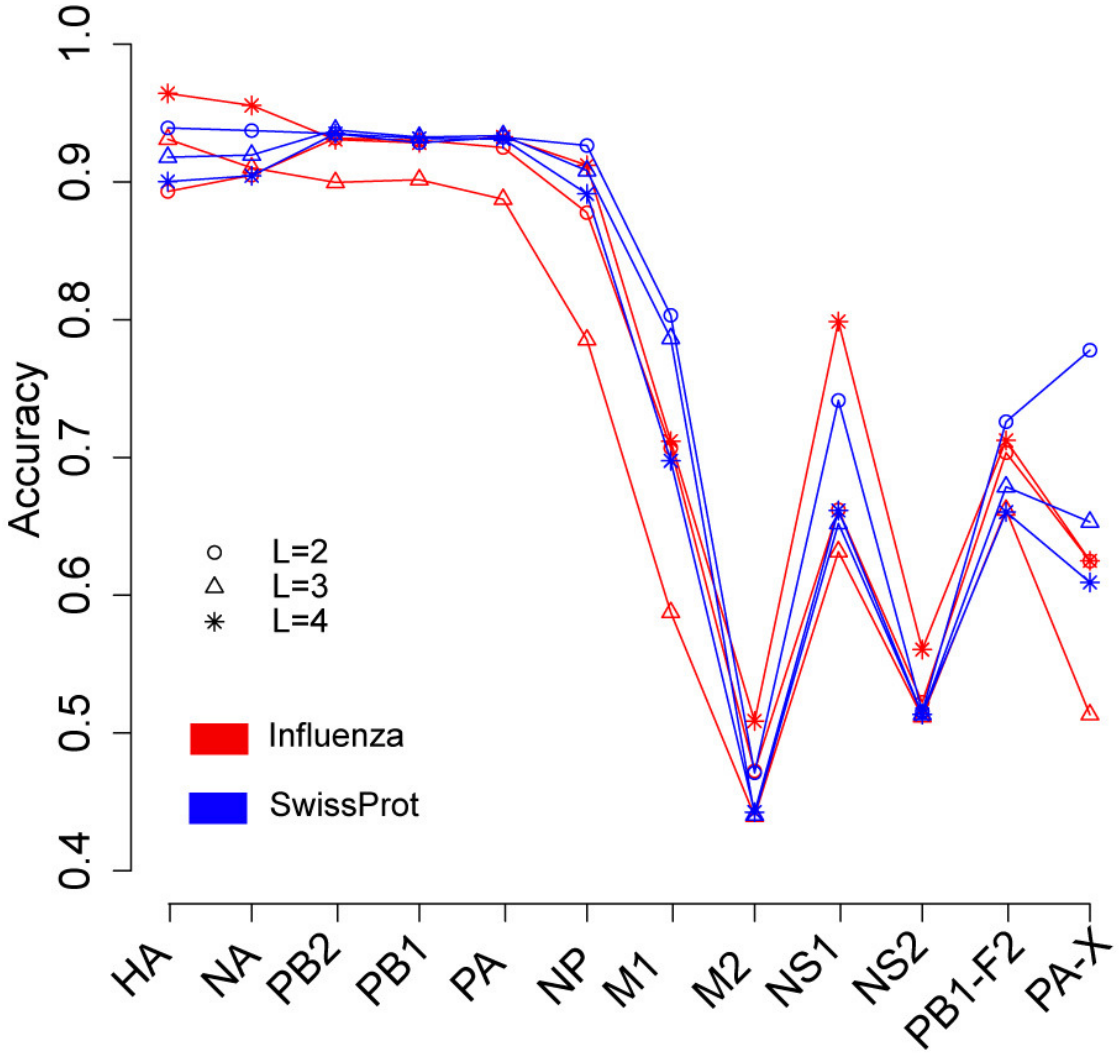


Figure 2 Performances for the models based on the word vector with words of two to four letters long (shown in circle, triangle and star, respectively) derived from the influenza protein dataset. In black the overall prediction accuracy is shown, while red, purple and blue lines represent the accuracies for avian, human and swine influenza viruses, respectively. The accuracies are averaged in ten-fold cross-validations.


Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
L=2
L=3
L=4
Overall accuracy
Accuracy_avian
Accuracy_human
Accuracy_swine
HA
NA
PB2
PB1
PA
NP
M1
M2
NS1
NS2
PB1-F2
PA-X

335

336 **Figure 3** Comparison of overall accuracies for the models based on word vectors
337 with words of two to four letters long (shown in circle, triangle and star,
338 respectively) derived from the influenza protein dataset (red line) and SwissProt
339 dataset (blue line). The accuracies were averaged in ten-fold cross-validations.

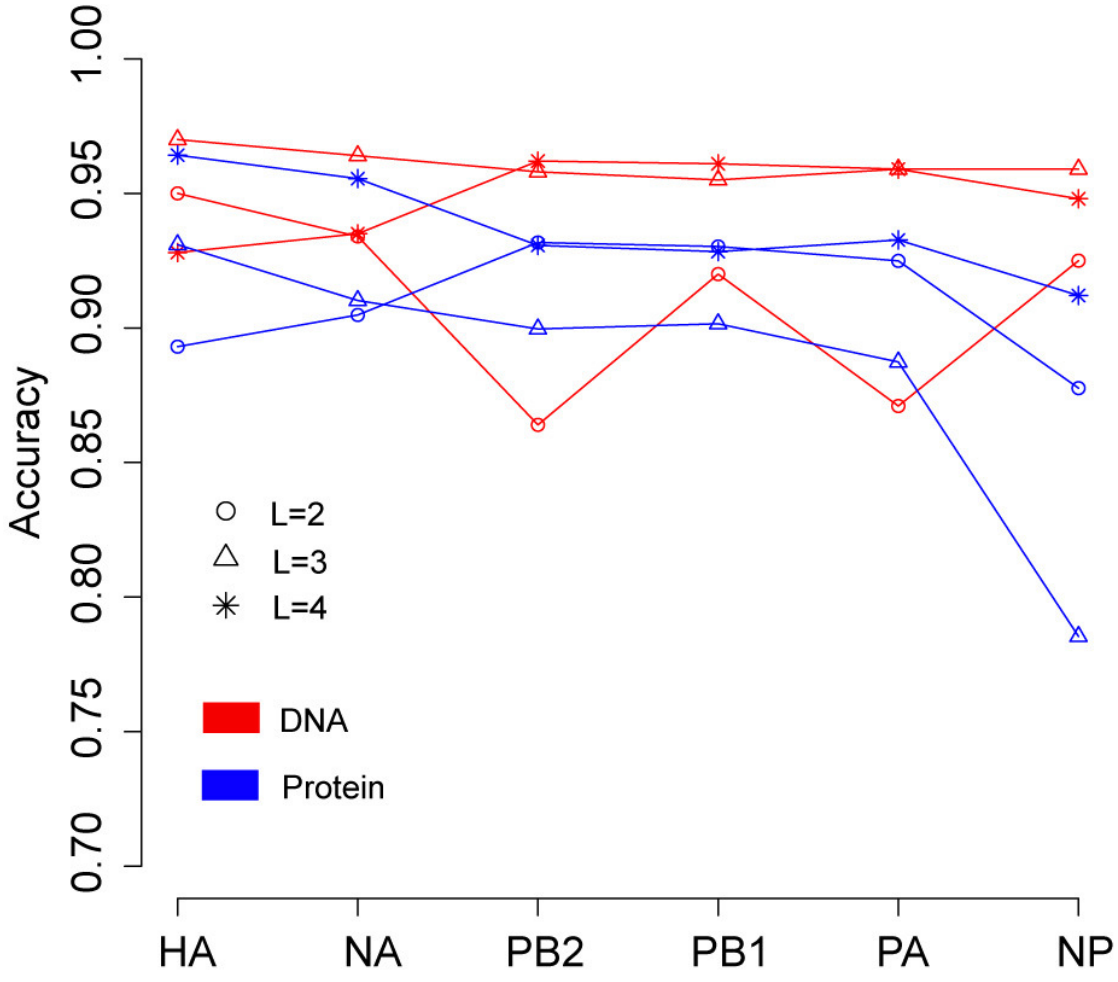


340

341

342

343 **Figure 4** Comparison of overall accuracies for the models based on word vectors
344 with words of two to four letters long (shown in circle, triangle and star,
345 respectively) derived from the influenza DNA dataset (red line) and influenza
346 protein dataset (blue line). The accuracies are averaged in ten-fold cross-
347 validations.



348