35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Signal-processing-based bioinformatics approach for the identification of influenza A virus subtypes in Neuraminidase genes

Charalambos Chrysostomou[1]* and Huseyin Seker[2]

Abstract— **Neuraminidase (NA) genes of influenza A virus is a highly potential candidate for antiviral drug development that can only be realized through true identification of its sub-types. In this paper, in order to accurately detect the sub-types, a hybrid predictive model is therefore developed and tested over proteins obtained from the four subtypes of the influenza A virus, namely, H1N1, H2N2, H3N2 and H5N1 that caused major pandemics in the twentieth century. The predictive model is built by the following four main steps; (i) decoding the protein sequences into numerical signals by means of EIIP amino acid scale, (ii) analysing these signals (protein sequences) by using Discrete Fourier Transform (DFT) and extracting DFT-based features, (iii) selecting more influential sub-set of the features by using the F-score statistical feature selection method, and finally (iv) building a predictive model on the feature sub-set by using support vector machine classifier. The protein sequences were chosen as to be of high percentage identity that they demonstrate within individual influenza subtype classes and high variation that they display in the percentage identity. This makes the proteins very difficult to distinguish from each other even they belong to different subtypes. Given this set of the proteins, the predictive model yielded 98.3% accuracy based on a 5-fold cross validation. This also results in a twenty feature sub-set that can also help reveal spectral characteristics of the subtypes. The proposed model is promising and can easily be generalized for other similar studies.**

Index Terms— **Amino Acid Indices, Discrete Fourier Transform (DFT), F-score, Neuraminidase Genes, Support Vector Machines**

I. Introduction

In recent years, decoding the rules that drive biological functions of influenza subtypes directly from their primary structures, has become a subject of intensive research. Signal processing-based techniques such as Resonant Recognition Model (RRM) [1]–[3] and Complex Resonant Recognition Model (CRRM) [4] have been introduced in bioinformatics to extract information that is expected to match protein biological functions. The study is performed using the algorithms that help derive meaningful knowledge from the proteins based on features extracted from the signal processing techniques.

For this study different neuraminidase (NA) genes of influenza A virus subtypes are selected and presented including H1N1, H2N2, H3N2 and H5N1 NA subtypes. These protein sequences were chosen for the high percentage identity they demonstrate within individual influenza subtype classes and the high variation they display in percent identity.

In the literature various methods exist that can extract features directly from protein sequence primary structure with one example being the Basic Local Alignment Search Tool (BLAST) [5]. These methods perform very accurately for high homology sequences, whereas their performance is considerably decreased for low homology sequences. Therefore, a new homology independent method is needed to be developed in order to extract features from the primary sequence structure and to be able to identify all the important features that can be related to the bioinformatics problem, and to discard any ineffective or noisy data.

Signal processing techniques can generate a large amount of information, which can be related to a protein's biological function. The RRM and CRRM are only two of the techniques that try to identify which of the features extracted are related to the protein's biological function. In this study, F-Score and Support Vector Machine (SVM) [6], [7] are utilized to be able to determine if a feature, or a set of features, extracted from protein sequences using signal processing techniques can be used to characterise different protein classes.

In this paper, SVM is implemented to create a classification model, which can be used to model relationships between protein sequences. SVM is a supervised statistical learning method that analyses data and recognises patterns for classification [6], [7]. The SVM takes a set of input data and predicts to which of two or more possible classes each given input protein belongs. SVM is selected as it can produce accurate and robust classification results on a established theoretical basis even when input data are noisy or non linearly separable [8], [9]. The predictive model obtained by SVM with the complete data set is presented to show the more representative subgroups and classification models created for the influenza A virus problem.

The paper is organised as follows: Section II presents the methods and materials used in this paper including the protein sequences that belong to the NA genes (Section II-A), the signal processing method, namely Discrete Fourier Transform, used to extract protein related features (Section II-B), Feature selection using F-score (Section II-C) and SVM-based classifier (Section II-D). Section III presents a case study with influenza subtypes sequences and the results obtained by SVM. Finally, conclusions are discussed in Section IV.

[1]Department of Genetics, University of Leicester, University Road Leicester, LE1 7RH, United Kingdom
[2]Bio-Health Informatics Research Group, Centre for Computational Intelligence, Faculty of Technology, De Montfort University, Leicester, LE1 9BH, UK
charalambos.chrysostomou@gmail.com, hseker@dmu.ac.uk
*Corresponding Author

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

II. Methods and Materials

A. Protein sequences for influenza A virus subtypes in Neuraminidase genes

Influenza A virus belongs to the orthomyxoviridae family of viruses and can affect mainly birds and some mammals. The Influenza A virus genome consists of eight single genes; the hemagglutinin (HA) gene, the neuraminidase (NA) gene, the nucleoprotein (NP) gene, the matrix proteins (M) gene, the non-structural proteins (NS) gene and three RNA polymerase (PA, PB1, PB2) genes. Human pandemics outbreaks rarely arise when the influenza A virus is transmitted from wild birds to domestic poultry. During the twentieth century, three major influenza pandemics were recorded, which were caused by H1N1, H2N2, and H3N2 viruses. In addition, the H5N1 virus is considered as a current pandemic thread. For this analysis, as Table I shows, four different subtypes of Influenza A virus Neuraminidase (NA) gene were used, as it is the target for current antiviral drugs, called neuraminidase inhibitors [10].

TABLE I
INFLUENZA A VIRUS NEURAMINIDASE PROTEINS

Subtype	No of Sequences	Period
H1N1	200	2009
H2N2	76	1957-1968
H3N2	200	1968-2000
H5N1	70	2005-2009

For influenza A subtypes 200 H1N1 NA proteins from 2009, 76 H2N2 NA proteins from the period 1957-1968, 200 H3N2 NA proteins from the period 1968-2000 and 70 H5N1 NA proteins from the period 2005-2009 were collected from the Influenza Virus Resource data set [11]. The relationship of influenza subtypes in respect of NA gene is shown in the following:

- H1N1 from 2009 is the result of reassortment between Eurasian H1N1 influenza A swine virus and H1N2 swine virus [12]. H1N1 retains the NA gene from Eurasian H1N1 influenza A swine virus.
- H2N2 from the period 1957-1968 is the result of reassortment between existing human H1N1 and avian H2N2 viruses [12]. H2N2 retains the NA gene from the avian H2N2 virus.
- H3N2 from the period 1968-2000 is the result of reassortment between circulating human H2N2 and avian H3 viruses [12]. H3N2 retains the NA gene from human H2N2 virus.
- H5N1 from the period 2005-2009 was created by combining various influenza A subtype viruses [13] where H5N1 retains the NA gene from avian H1N1 virus.

Percentage identity is a measurement used to determine the similarity between protein sequences. By using CLUSTALW [12], the pairwise percent identity of all the influenza A NA genes was calculated. Table II shows the average percent identity between all the classes.

As Table II shows, the percent identity within each individual influenza subtype class is very high yielding 93%, 96%, 94% and 96% for H1N1 NA, H2N2 NA, H3N2 NA and H5N1 NA influenza A subtypes. In contrast to the individual class, percent identity from different classes may vary significantly, with high average percent identity of 83% between H1N1 and H5N1 and 86% between H2N2 and H3N2. Very low average percent identity was determined between H1N1 and H2N2 with 42%, H1N1 and H3N2 with 40%, H5N1 and H2N2 with 43%, and finally H5N1 and H3N2 with 41% average percent identity. The pairwise percent identity results presented in Table II suggest that two subtype pairs (H1N1 and H5N1) and (H2N2 and H3N2) contain highly similar proteins meaning that it is much more difficult to distinguish the proteins from each other compared to other subtype pairs.

TABLE II
AVERAGE PAIRWISE PERCENT IDENTITY

	H1N1	H2N2	H3N2	H5N1
H1N1	93%	-	-	-
H2N2	42%	96%	-	-
H3N2	40%	86%	94%	-
H5N1	83%	43%	41%	96%

B. Signal Processing For Protein Sequence Analysis

By using digital signal processing techniques, the goal is to extract information that can be related to biological functions of proteins. Various methods have been used in bioinformatics for analysing protein sequences in recent years, and one of the most common methods is the RRM [1]–[3] and CRRM [4]. Previous studies [14] used influenza A subtypes to analyse the hemagglutinin (HA) gene, with RRM aiming to identify new therapeutic targets for drug development by better understanding the interaction of the influenza virus and its receptors.

In contrast to previous studies, the analysis was performed directly to absolute spectrum, which is derived by applying Discrete Fourier Transform (DFT) to each numerical encoded protein sequence. Electron-ion interaction potential (EIIP) [15], [16] amino acid index, as shown in Table III, is used to turn protein sequences into numerical sequences in order to be able to apply DFT. For the analysis of influenza A virus proteins, as the sequences have different lengths, zero-padding was used to extend all the protein sequences to $N = 512$ thus the output of the absolute spectrum contains 256 features.

TABLE III
EIIP VALUES

Amino acid	EIIP	Amino acid	EIIP
Leu	0.0000	Tyr	0.0516
Ile	0.0000	Trp	0.0548
Asn	0.0036	Gln	0.0761
Gly	0.0050	Met	0.0823
Glu	0.0057	Ser	0.0829
Val	0.0058	Cys	0.0829
Pro	0.0198	Thr	0.0941
His	0.0242	Phe	0.0946
Lys	0.0371	Arg	0.0959
Ala	0.0373	Asp	0.1263

C. Feature Selection Using F-score

Feature selection [17], [18] is the technique of selecting relevant features for building robust classification models. Furthermore, feature selection is a particularly important step in analysing the data from many experimental techniques as they often include a large number of variables but low number of samples. By removing redundant features from the data, feature selection can improve the performance of classification techniques like SVM in the following ways:

- Reduce data dimensionality.
- Improve the generalisation capability of the classification model.
- Speed up learning process.
- Improve model interpretability.

F-score is one of the simplest but effective techniques that measures the separation of two sets of real numbers [19].

D. Support Vector Machines

A support vector machine, (SVM) [6], [7] is a supervised statistical learning method that analyses data and recognises patterns for classification. The SVM takes a set of input data and predicts to which of two possible classes each given input belongs. SVM is used in this analysis as it can produce accurate and robust classification results on a established theoretical basis even when input data are noisy or non linearly separable [8], [9]. For this analysis the LIBSVM [20] tool was used to build a classification model. Furthermore, a 5-fold cross-validation was used in combination with F-score to find the optimum number of features that can be used to predict Influenza A neuraminidase subtypes without sacrificing any accuracy. In addition, grid search was used to find the optimal SVM parameters for the predictive model.

III. Results and Discussions

By using SVM and F-score, a classification model was constructed for the Influenza A neuraminidase gene subtypes. By using F-score the goal is to select the most separable features extracted from influenza subtypes and create a predictive model without sacrificing any of the accuracy obtained by using all the features extracted. Figure 1 shows F-score value for all the features extracted from the protein sequences.

In order to build an accurate and generalised predictive model, 5-fold cross-validation was used. In combination with F-score, the minimum number of useful features that can be used to predict Influenza A neuraminidase subtypes without sacrificing any accuracy was found to be 20. This number of features was discovered by manually eliminating features with the lowest F-score and repeating the analysis. Table IV shows the best 20 stratified spectral characteristic features along with their F-score values.

The total accuracy for the complete data set is 0.983 ± 0.006. As the results show, good precision in classifying new protein sequences can be obtained. An analysis for each influenza subtype can be observed below:

- For the H1N1 subtype class the average accuracy is 1.0 ± 0.0.

TABLE IV
Top 20 Features In Order of Importance Based on F-score

	Feature	Score		Feature	Score
1	39	25.0333	11	116	12.1021
2	9	20.8762	12	136	11.8541
3	97	20.8564	13	79	11.6046
4	8	20.7845	14	90	11.4991
5	38	15.2451	15	192	10.3009
6	98	15.0310	16	117	9.9620
7	209	13.9315	17	234	9.8864
8	7	13.1166	18	252	9.1989
9	197	12.9016	19	236	9.1070
10	111	12.6627	20	113	8.9069

TABLE V
Confusion Matrix For SVM Predictive Model

Class	H1N1	H2N2	H3N2	H5N1
H1N1	**1.00 ± 0.00**	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
H2N2	0.00 ± 0.00	**0.92 ± 0.05**	0.08 ± 0.05	0.00 ± 0.00
H3N2	0.00 ± 0.00	0.01 ± 0.01	**0.99 ± 0.01**	0.00 ± 0.00
H5N1	0.04 ± 0.04	0.00 ± 0.00	0.00 ± 0.00	**0.96 ± 0.04**

- For the H2N2 subtype class the average accuracy is 0.92 $\pm$ 0.05, where the misclassified proteins were as H3N2.
- For the H3N2 subtype class the average accuracy is 0.99 $\pm$ 0.01, where the misclassified proteins were as H2N2.
- For the H5N1 subtype class the average accuracy is 0.96 $\pm$ 0.04, where the misclassified proteins were as H1N1.

This analysis shows a strong correlation between the features extracted and refined using F-score with protein percentage identity between classes as shown in Table II. Subtypes that present low percentage identity between them, are classified with very high accuracy. The most challenging part is to separate subtypes that present high percentage identity between them. As the bibliography indicates [12], [13], there is a clear biological connection between these influenza subtypes. Despite considerably much higher homology that exists in two subtype pairs (H1N1 and H5N1) and (H2N2 and H3N2), the predictive model seems to have overcome this problem yielding near perfect predictive accuracy of 96% and 99%, respectively.

IV. Conclusions

The paper presents a highly successful predictive model with an accuracy of 98.3% that has helped distinguish the four subtypes (H1N1, H2N2, H3N2 and H5N1) that belong to the Neuraminidase (NA) genes of influenza A virus that has recently been regarded as highly potential antiviral drug candidate. It is particularly worth noting that although considerably much higher homology is observed in two subtype pairs (H1N1 and H5N1) and (H2N2 and H3N2), the predictive model seems to have overcome this problem yielding near perfect predictive accuracy of 96% and 99%, respectively. In addition, it has been demonstrated that the signal processing technique, namely Discrete Fourier Transform, was found to generate useful spectral characteristic features that are highly capable of representing the protein groups and that this was further enhanced using the F-score feature selection method and SVM-based classifier.

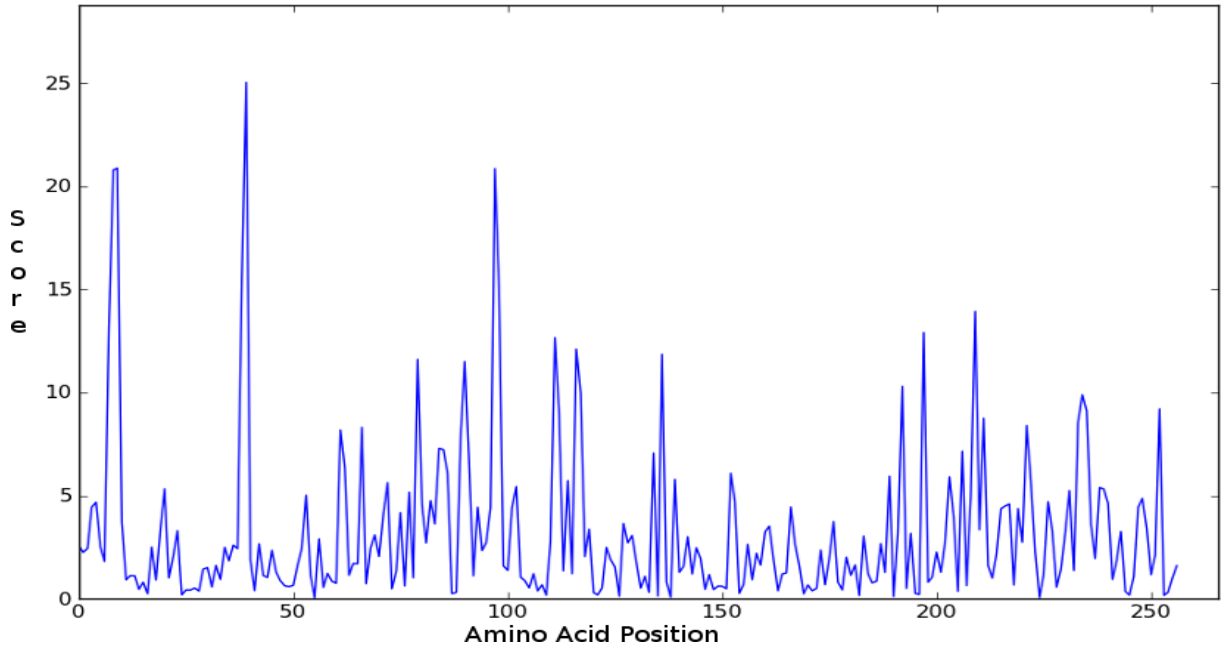

Fig. 1. Feature Scores Based on F-score

In addition to the EIIP amino acid index (Table III) used in this study, there are over 500 amino acid indices reported in the literature [21] that can represent different biological features. They can therefore be used to construct different models in future studies in order to identify which of these amino acid scales is more representative of these subtypes as well as others. This is required as to monitor future outbreaks more accurately and identify better drug candidates, which can only be realized through true identification of the subtypes.

REFERENCES

[1] E. Pirogova, "Analysis of amino acid parameters in the resonant recognition model," *Proceedings of the International Conference on Bioelectromagnetism*, p. 71, 1998.

[2] E. Pirogova and I. Cosic, "Bioactive peptide design using the resonant recognition model," *Nonlinear Biomed Phys.*, p. 17, 2007.

[3] I. Cosic, "Macromolecular bioactivity: is it resonant interaction between macromolecules?Theory and applications." *IEEE transactions on bio-medical engineering.*, vol. 41, p. 1101, 1994.

[4] C. Chrysostomou, H. Seker, N. Aydin, and P. Haris, "Complex resonant recognition model in analysing influenza a virus subtype protein sequences," in *10th IEEE International Conference on Information Technology and Applications in Biomedicine*, Corfu, Greece, November 2010, pp. 1–4. [Online]. Available: http://dx.doi.org/10.1109/ITAB.2010.5687621

[5] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[6] B. E. Boser and et al., "A training algorithm for optimal margin classifiers," in *PROCEEDINGS OF THE 5TH ANNUAL ACM WORKSHOP ON COMPUTATIONAL LEARNING THEORY*. ACM Press, 1992, pp. 144–152.

[7] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[8] M. Hearst, S. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *Intelligent Systems and their Applications, IEEE*, vol. 13, no. 4, pp. 18–28, 1998.

[9] I. Steinwart and A. Christmann, *Support vector machines*. Springer Verlag, 2008.

[10] A. Moscona, "Neuraminidase inhibitors for influenza," *New England Journal of Medicine*, vol. 353, no. 13, p. 1363, 2005.

[11] Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman, "The influenza virus resource at the National Center for Biotechnology Information," *Journal of virology*, vol. 82, no. 2, p. 596, 2008.

[12] D. Morens, J. Taubenberger, and A. Fauci, "The persistent legacy of the 1918 influenza virus," *The New England journal of medicine*, vol. 361, no. 3, p. 225, 2009.

[13] M. M. Mukhtar, S. T. Rasool, D. Song, C. Zhu, Q. Hao, Y. Zhu, and J. Wu, "Origin of highly pathogenic H5N1 avian influenza virus in China and genetic characterization of donor and recipient viruses," *JOURNAL OF GENERAL VIROLOGY*, vol. 88, no. Part 11, pp. 3094–3099, NOV 2007.

[14] V. Veljkovic, N. Veljkovic, C. Muller, S. Muller, S. Glisic, V. Perovic, and H. Kohler, "Characterization of conserved properties of hemagglutinin of h5n1 and human influenza viruses: possible consequences for therapy and infection control," *BMC Structural Biology*, vol. 9, no. 1, p. 21, 2009.

[15] V. Veljkovic, I. Cosic, B. Dimitrijevic, and D. LalovicC, "Is it possible to analyze DNA and protein sequences by the methods of digital signal processing?" *IEEE Transaction on Biomedical Engineering*, vol. 32, no. 5, pp. 337–341, 1985.

[16] K. Gopalakrishnan, R. Zadeh, K. Najarian, and A. Darvish, "Computational analysis and classification of p53 mutants according to primary structure," in *2004 IEEE Computational Systems Bioinformatics Conference, Proceedings*, 2004, Proceedings Paper, pp. 694–695.

[17] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 1-4, pp. 131–156, 1997.

[18] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Springer, 1998, vol. 454.

[19] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*. Springer, 2006, vol. 207.

[20] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[21] S. Kawashima, H. Ogata, and M. Kanehisa, "AAindex: amino acid index database," *Nucleic Acids Research*, vol. 27, no. 1, p. 368, 1999.