Application of bioinformatics models to define influenza virus A subtypes

M. Ebrahimi[1], P. Agha-Golzadeh[2], E. Ebrahimie[2] and N. Shamabadi[3]
[1]Department of Biology & Bioinformatics Research Group, University of Qom, Qom, Iran
[2]Department of Crop Production & Plant Breeding, College of Agriculture, Shiraz University, Shiraz, Iran
[3]Young Researcher Club, Qom Branch, Islamic Azad University, Qom, Iran

Abstract - *Influenza A viruses infect large numbers of animals and are subtyped according to their surface antigens to 16 HA subtypes and 9 NA subtypes. To identify the main prominent protein attributes representing each subtype, various clustering, screening, item set mining and decision tree models applied to dataset of 3632 HA sequences of influenza A viruses. The count of Tyr, Gln and Phe and the count of some hydrophilic – hydrophobic (such as Lys – Val, Asn – Leu and Pro – Leu) were the most important protein features. Most decision tree models used non-reduced absorption at 280nm as the main protein feature to build the trees. Parallel stump and ID3 numeric decision tree algorithms were the best tree to differentiate between HA subtypes. The results showed various bioinformatics tools may be used in this regard. For the first time, this paper showed that protein attributes can be used to differentiate between influenza A subtypes.*

Keywords: Influenza A, Bioinformatics, Modelling

1 Introduction

Influenza is a highly contagious and acute respiratory disease with a high degree of morbidity and has been in circulation for centuries [1]. The disease is caused by the influenza virus, which is a segmented, enveloped RNA virus. Within the influenza virus family, there are four genera: A, B, C virus and Thogoto virus; although only A and B cause significant disease in humans [2]. Influenza A viruses are further subtyped according to their surface antigens, HA and NA, of which 16 HA subtypes and 9 NA subtypes have been identified to date [3]. The HA and NA genes are extremely variable in sequence, and less than 30% of the amino acids are conserved among all the subtypes. New epidemic strains of influenza A arise due to point mutations within two surface glycoproteins, HA and NA. These changes in HA and NA enable emerging virus strains to evade the host's immune system and therefore necessitates the annual revision of vaccine to include the new viruses [4]. Furthermore, HA may also play a structural role in budding and particle formation. Human influenza viruses manage to cause epidemics almost every year. The circulating viruses change their surface glycoproteins by accumulating mutations (antigenic drift or antigenic shift) which results in variant viruses of the same subtype that are able to evade the immune pressure in the population [5].

Bioinformatics represents a new field at the interface of the twentieth-century revolutions in molecular biology and computers. A focus of this new discipline is the use of computer databases and computer algorithms to analyze proteins and genes. A major challenge in biology is to make sense of the enormous quantities of sequence data and structural data that are generated by genome-sequencing projects, proteomics, and other large-scale molecular biology efforts. Fitting a model such as a decision tree or item set mining to a set of variables this large may require more time than is practical [6]. A decision tree is constructed by looking for regularities in data, determining the features to add at the next level of the tree using an entropy calculation, and then choosing the feature that minimizes the entropy impurity [7]. To better understand the features that contribute to structural differences between influenza viruses A subtypes, it is necessary to identify the main features responsible for this valuable characteristic. Herein we used various clustering, screening, item set mining and decision tree models to determine which protein attributes may be used as a marker between subtypes of influenza A viruses. All available HA sequences (3632) of influenza A viruses from Swiss-Prot database were extracted and up to 924 protein features for each HA protein sequence was generated and various bioinformatics modeling techniques applied on this.

2 Methods and Materials

Three thousand and six hundred and thirty two sequences of HA virus proteins from various species (human, bird, pig, horse, mouse, tiger, leopard, dog, and cat) were extracted from the UniProt knowledgebase database and categorized as H1 to H16, according to database classification. Nine hundred and twenty four protein features or attributes including primary and secondary protein features were extracted. A dataset of these protein features was imported into Clementine software [Clementine_NLV-11.1.0.95; Integral Solution, Ltd.], null data for subtype of virus was discarded, and this feature was set as the output variable and the other variables were set as input variables. The same database imported into RapidMiner software [RapidMiner 5.0.001, Rapid-I GmbH, Stochumer Str. 475, 44227 Dortmund, Germany] and again the subtype of virus set as target or label attribute [when Item

Set Mining model performed, no label or target attribute was set as this model requires so]. To minimize the effects of correlated features on modelling and to decrease the processing time and burden on processing facilities, the original database subjected to remove correlated features algorithm, so the number of protein attributes (variables) decreased from 924 to 486 attributes. Various algorithms such as screening models [Anomaly detection model, feature selection algorithm or attribute weighting], clustering models [K-Means, TwoStep cluster], Tree Induction models [with various criterion, C5.0, C5.0 with 10-fold cross Validation and C&RT], Item Set Mining [FPGrowth] and Rule Induction model [10 fold cross-Validation through stratified sampling] run on each dataset as described previously [8]. Whenever requested by model, data were discretized by the frequency; i.e. data were divided into 3 bins [ranges] with nearly equal the frequencies in each class [low 0-0.3, mid 0.3-0.5 and high >0.5]; and sometimes data were converted to nominal and in some cases to binominal datasets.

3 Results

The number of protein attributes gained weights higher than 0.7 in each weighting model were as follows: PCA 2, SVM 24, relief 4, uncertainty 17, gini index 280, chi squared 39, deviation 2, rule 59, gain ratio 61, info gain 350 and info gain ratio 13.

The most important feature used to build the tree was non-reduced absorption at 280nm. If the value for this protein attribute was higher than 1.180 and the value for the count of Trp – Ala was higher than 0.500 and the count of Gly was higher than 49, the viral protein was originated from H10; otherwise from H3. If the count of Trp – Ala was equal to or less than 0.500, then the count of Ala – Ala (value of 3.500), the length of protein (value of 566) and the count of Trp – Asn (value of 0.500) used to differentiate between H14, H4, H8 and H9 groups. When the count of Trp – Asn was higher equal to or less than 0.500, if the count of Ser – Cys, non – reduced absorption at 280nm and aliphatic index were higher than 1.500, 1.44 and 86.690, respectively, the protein originated from H16; otherwise from H13. With the count of Ser – Cys was equal to or less than 1.500 and the count of His – Asn was higher than 0.500 and the count of Glu – Trp was higher than 0.500, if the count of Gly was higher than 44.500, virus belonged to H2, otherwise to H5 group. With the count of Glu – Trp (< 0.500) and the count of His – Asn (<.500), the virus HA proteins belonged to H1 and H6, respectively. If non-reduced absorption at 280nm was < 1.180 and the aliphatic index was > 81.875, the protein belonged to H12 group, if not to H15 or H7.

Stump decision tree model created a very simple tree with non-reduced absorption at 280nm variable as the root feature. Decision Tree Stump (Parallel) generated a tree again with the same starting attribute. More complex tree generated by ID3 Numerical method and again tree built on non-reduced absorption attribute. Random tree started with another protein attribute, the count of His – Ala. When value for this attribute was higher than 1.500 and the count of Ala was higher than 26.500, the protein fell into H6 group. if the count of His-Ala was higher than 1.500 and the count of Ala was less than or equal to 26.500, the virus protein identified as H16. Ten different models were used by Random Forest algorithm to induce decision trees. In the first model, the count of Met-Ala was the main feature used by this method to induce the tree and its branches was created using the count of Gly and the count of Vla – Arg attributes to classify H2, H5, H10, H9, H8, H7, H1 and H11 subtypes. In the second model, the count of Gly – Ala, the frequency of Pro – Ile, the count of Asn – Cys, the frequency of Pro – Ile, the count of Met – Lys and the count of Leu – Trp to trace H6, H11,H1, H3, H5, H13, H2 and H9 subtypes. The count of Gly – Met, the count of Cys – His and the count of sulfur were the most important attributes to build the tree by the third model (H10, H3, H9H4 and H5). Random forest, the fifth model, was able to differentiate between H10, H1, H4 and H3 by inducing a tree with the frequency of Pro – Ser as the main feature and the count of Cys – Met as the other important feature. In other models the count of Gln – Phe, the count of Trp – Pro and the count of Ala – Ala (model 5), the count of His – Phe, the count of Ile – Phe, the count of Leu – Lys and the count of Ala – Gln (model 6), the count of Gln – Gln and the count of Gln – Tyr (model 7), the count of Phe – Lys, the count of Asn – His and the count of Ser – Pro (model 8), the count of Gly – Met, the count of Gly – Val, the count of Asp – Gly and the count of Pro – Ala (model 9) and the count of Trp – Met (model 10) were the most important features used to build the trees.

GRI node analysis created 100 rules with 3631 valid transactions with minimum and maximum support of 44.09 % and 44.09 %, respectively, while maximum confidence reached 100 %. When feature selection was used, minimum support, maximum support, maximum confidence, and minimum confidence were the same as previous. In both methods [with/without feature selection filtering] the count of Gln – Leu, the frequency of Gly – His and the frequency of Pro – Asn were the main features used to create the first rules.

4 Discussion

Although the numbers of attributes with weights equal to or higher than 0.70 varied from 2 (in PCA weighting) to 62 (in Info Gain Ratio and Rule Induction weighting), the percentage and the count of Tyr, the frequency and the count of Lys - Val, the percentage, the frequency and the count of Gln, the frequency and the count of Asn – Leu, the count of Pro – Leu, the percentage of Phe and the frequency of Ser – Ile chosen by 7 attribute weightings as one of the most important attributes. When the same models run on dataset with correlated remove features, only six attributes gained weights higher than 0.70; again the count of Tyr, the count of Gln, the count of Lys – Val and the count of Asn – Leu were the most important features with weights higher than 0.70. The count of Gln – Asn was the other weight higher than 0.70. More than 50% of features gained high weights in both models were hydrophobic amino acids and the rest were mainly from hydrophilic amino acids. For the first time the importance of dipeptides in classifying the influenza virus A has been presented here. The combination of one hydrophobic amino acid such as Val, Leu or Ile and one hydrophilic amino acid such as Asn, Ser or Gln forms a strong link inside the protein and reduce the possibility of mutations in this area; but when there are hydrophobic dipeptides connections, the chance of mutation and flexibility increases.

Although some trees generated by tree induction models had just two branches, as seen in stump decision tree, the depth of trees in some models were more complicated [more than 12 branches in ID3 numeric run on information gain]. The ability of various decision tree induction models applied in this study to correctly and effectively classify influenza A subtypes based on protein attributes were very different. In some models, two or three classes were identified, showing the model was not competence in this field (as seen in decision tree stump, C5.0, C&RT, random tree and accuracy). But in some other models, such as decision tree run on removed correlated features' dataset, decision tree stump parallel and ID3 numeric, the models were able to completely classify the HA subtypes (H1 – H16) based on their protein features. So the latter models may be used as a suitable tool to classify those viral subtypes.

The results showed that various bioinformatics tools and modelling facilities can be used to identify the subtypes of influenza virus A with a precision rate up to 95%. To our knowledge, for the first time we showed that some primary or secondary attributes can be used to differentiate between various subtypes of influenza A viruses.

5 References

[1] P. Chadha and R. H. Das, "A pathogenesis related protein, AhPR10 from peanut: an insight of its mode of antifungal activity," *Planta,* vol. 225, pp. 213-22, Dec 2006.

[2] A. M. Ledeboer, *et al.*, "Cloning of the natural gene for the sweet-tasting plant protein thaumatin," *Gene,* vol. 30, pp. 23-32, Oct 1984.

[3] C. Kuwabara, *et al.*, "Abscisic acid- and cold-induced thaumatin-like protein in winter wheat has an antifungal activity against snow mould, Microdochium nivale," *Physiol Plant,* vol. 115, pp. 101-110, May 2002.

[4] H. Breiteneder and C. Ebner, "Molecular and biochemical classification of plant-derived food allergens," *J Allergy Clin Immunol,* vol. 106, pp. 27-36, Jul 2000.

[5] Y. Tada and T. Kashimura, "Proteomic analysis of salt-responsive proteins in the mangrove plant, Bruguiera gymnorhiza," *Plant Cell Physiol,* vol. 50, pp. 439-46, Mar 2009.

[6] A. M. Casas, *et al.*, "Expression of Osmotin-Like Genes in the Halophyte Atriplex nummularia L," *Plant Physiol,* vol. 99, pp. 329-37, May 1992.

[7] D. Goel, *et al.*, "Overexpression of osmotin gene confers tolerance to salt and drought stresses in transgenic tomato (Solanum lycopersicum L.)," *Protoplasma,* vol. 245, pp. 133-41, Sep 2010.

[8] A. Leone, *et al.*, "Comparative Analysis of Short- and Long-Term Changes in Gene Expression Caused by Low Water Potential in Potato (Solanum tuberosum) Cell-Suspension Cultures," *Plant Physiol,* vol. 106, pp. 703-712, Oct 1994.

[9] M. Ebrahimi and E. Ebrahimie, "Sequence-based prediction of enzyme thermostability through bioinformatics algorithms," *Current Bioinformatics,* vol. 5, pp. 195-203, 2010.

[10] M. Ebrahimi, *et al.*, "Searching for patterns of thermostability in proteins and defining the main features contributing to enzyme thermostability through screening, clustering, and decision tree alogorithms," *EXCLI Journal,* vol. 8, pp. 218-233, 2009.

[11] M. Ebrahimi, *et al.*, "Are there any differences between features of proteins expressed in malignant and benign breast cancers?," *Journal of Research in Medical Sciences,* vol. 15, pp. 299-309, 2010.

[12] E. Ebrahimie, *et al.*, "Investigating protein features contribute to salt stability of halolysin proteins," *Journal of Cell and Molecular Research,* vol. 2, pp. 15-28, 2010.

[13] E. Ashrafi, *et al.*, "Determining specific amino acid features in P1B-ATPase heavy metals transporters which provides a unique ability in small number of organisms to cope with heavy metal pollution " *Bioinformatics and Biology Insights,* vol. Accepted, 2011.

[14] M. Ebrahimi, *et al.*, "Searching for patterns of thermostability in proteins and defining the main features contributing to enzyme thermostability through screening, clustering, and decision tree algorithms," *EXCLI Journal,* vol. 8, pp. 218-233, 2009.