Text Mining as a Strategy in Profiling the Use of Influenza Virus Genome in Scientific Publications

Fernanda C. R. Correa
Federal University of Health Sciences of Porto Alegre (UFCSPA)
Porto Alegre, Brazil
Email: fernandacr [AT] ufcspa.edu.br

Aline A. Vanin
Federal University of Health Sciences of Porto Alegre (UFCSPA)
Porto Alegre, Brazil

Silvio C. Cazella
Federal University of Health Sciences of Porto Alegre (UFCSPA)
Porto Alegre, Brazil

Abstract — **The aim of this study was to profile the use and usage patterns of influenza virus genome from scientific publications in online databases using Natural Language Processing and Text Mining techniques. A systematic research was performed to select papers in PubMed electronic database using the keywords: 'influenza', 'genome', 'database'. The 45 articles that presented free full text available were processed with the sofwares AntFileConverter and AntConc. Text Mining was performed with the software Weka. Association rules were expected between genome and influenza. Also, it was predicted that influenza genome and terms related directly to the application of genome databases would relate. However, the results revealed an association between influenza virus protein and mutation sequence/database. The discovery of different associations than the expected revealed the necessity of expanding the research in order to increase the size of the corpus and to improve the attributes selection for mining in Weka sofware.**

Keywords – Data Mining; Natural Language Processing; Influenza A virus; Genome, Viral; Databases, Nucleic Acid

I. INTRODUCTION

New approaches based upon molecular and computational methods are essential for advances in the study and control of infectious diseases. In this context, pathogenic viruses such as influenza viruses are an important source of study for the development of these new methods, given the large amount of information available on these microorganisms [1].

Influenza virus is an important human pathogen that causes a high number of deaths every year. Seasonal influenza epidemics result in over three million cases of severe illness, and about 250,000 to 500,000 deaths every year [2]. Influenza virus has an annual attack rate estimated at 5-10% in adults and 20-30% in children, with possible hospitalization and death specially in more susceptible population such as the elderly, the chronically ill, and pregnant women [3].

Prevention and control of influenza epidemics is a major problem for public health care services [4]. The current approach is annual vaccination with a trivalent inactivated influenza vaccine, composed by two different influenza type A strains and one influenza type B strain [5].

Influenza A virus has a high mutation rate and wide host range. Avian H5N1 and H7N9 were able to cause human infections, raising the fear of a new influenza strain that could result in a global pandemic due to the absence of previous host immunity [6] [7]. However, due to constant epidemiological control by health authorities, there is more control of a possible outbreak. [8] [9] [10].

Biological data from online databases provide an excellent source of material for research. Genomes sequences are easily and quickly obtained [11]. Due to the recent advent of methodologies allowing the analysis of these materials, mutations can be identified and used to predict the emergence of new strains with pathogenic potential as well as to understand how viruses spread geographically [12].

The data generated and accumulated in biological databases is consistent and abundant, creating an overload of information. Thus, computational techniques and new technologies are necessary to provide effective and efficient analysis of this content. Moreover, it is important to evaluate the amount of information that is generated and also to detect the potential of knowledge discovery that result from the application of these technologies [1].

Natural Language Processing (NLP) and Text Mining are two instruments that can be used in order to identify and extract relevant information from medical journals. The NLP consists on processing natural language texts by computer to access their meaning. Text Mining is a variation on a field called Data Mining which discovers and extracts knowledge from data, comprising activities such as information retrieval, information extraction and data mining to find associations among the pieces of information extracted from many different texts [13].

Text Mining, also known as Knowledge-Discovery in Texts (KDT), refers generally to the extraction of interesting and non-trivial information and knowledge from unstructured text. KDT combines extraction techniques, information retrieval, NLP and summarization of documents with data mining methods. Text Mining systems have been applied to the biological research area since the late 1990s and have considerably improved since then [14] [15].

Association rule mining is a technique used to discover relationships among a large set of variables in a data set. In association rules for text mining, the focus is to study relationships and implications among topics that are used to characterize a corpus, aiming to discover relevant association rules within a corpus such as the presence of a set of terms in an article implying the presence of another term [14].

The aim of this study was to analyze scientific publications available online to verify the use of influenza virus genome from online databases using NLP and Text Mining techniques. Moreover, associations or usage patterns of influenza genomes for different purposes are expected to be discovered. The results will provide useful information regarding the scientific publications generated from the study of influenza virus genome and thus displaying the potential use of these data in future studies and publications.

II. METHODS

For this systematic analysis, the PubMed electronic database was researched by using the following keywords: 'influenza', 'genome', 'database'. A total of 76 results were obtained, and 45 articles that presented free full text available (using PubMed filter) were selected. No other keywords or restrictions were used.

The .pdf files were converted into .txt files using the AntFileConverter 1.0.0 software [16]. The text files were cleaned, for removal of abstracts, titles, author informations and references. Next, a list of the 30 most common and relevant words was selected with the corpus analysis software AntConc 3.4.3 [17]. The texts were tokenized using binary representation in a Microsoft Office Excel 2010 spreadsheet. An Attribute-Relation File Format (ARFF) file with 30 attributes and 45 data instances was created. The presence of attributes was indicated as "1" (number ONE), while null values were indicated as "?" (question mark). A partial representation of the ARFF file is shown in Figure 1.

Text mining was performed with the software Weka 3.6 [18] in order to find association rules using the Apriori algorithm at minimum support of 0.6 and lift of 1.1. The software generated ten rules, as shown in Figure 2.

III. RESULTS AND DISCUSSION

The support of an itemset is defined as the proportion of transactions in the data set which contain the itemset. In this database, a set of items only appeared as a rule if it has occurred in 60% (0,6) of all transactions. The lift metrics is able to assess whether two items are positively or negatively independent, and also determines if two items are independent of each other. A minimum lif of 1.1 will only select rules in which the items are positively independent of each other [19].

The best rules showing the most frequently associated words are the rules 2, 5, 6 and 8, presenting higher conviction values than the others. Lift values are the same. Conviction, confidence and leverage values are very similar among this four rules. The elevated values of conviction indicate a higher independence of the foregoing item in relation to the consequent item. The rules are resemblant, indicating an association between influenza virus protein and mutation sequence/database. The most interesting rules are shown in Table 1.

```
@relation terms
@attribute Sequence {?,1}
@attribute Influenza {?,1}
@attribute Virus {?,1}
@attribute Gene {?,1}
@attribute Protein {?,1}
@attribute Database {?,1}
@attribute Genome {?,1}
@attribute Genbank {?,1}
@attribute Sequencing {?,1}
@attribute Pandemic {?,1}
@attribute Mutation {?,1}
@attribute Genotype {?,1}
@attribute Research {?,1}
@attribute Antigenic {?,1}
@attribute Pcr {?,1}
@attribute Vaccine {?,1}
@attribute Method {?,1}
@attribute Drug {?,1}
@attribute Antiviral {?,1}
@attribute Resistance {?,1}
@attribute Synthesis {?,1}
@attribute Assay {?,1}
@attribute Primer {?,1}
@attribute Openflu {?,1}
@attribute Array {?,1}
@attribute Culture {?,1}
@attribute Transcriptome {?,1}
@attribute Proteome {?,1}
@attribute Diagnostic {?,1}
@attribute Phylogenetic {?,1}

@data
%
% 45 instances
%
1,1,1,?,1,1,1,1,1,1,1,1,?,?,?,?,1,?,1,?,?,?,?,1,1,?,?,?,?
1,1,1,1,?,1,1,1,1,1,?,1,1,1,?,1,1,?,?,?,?,?,?,?,?,?,?,1,1
1,1,1,1,?,1,1,?,1,?,1,?,1,?,?,1,1,?,?,?,?,1,?,?,?,?,?,?,?
1,1,1,1,1,1,1,?,?,1,?,?,1,1,?,?,1,?,?,?,?,?,?,?,?,?,?,?,1
1,1,1,1,1,1,1,1,1,1,1,1,?,1,1,?,?,?,?,?,?,?,?,?,?,?,?,?,?
```

Figure 1. Partial ARRF File

Considering the amount of selected terms and the content of the papers, it was expected that association rules were

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 1 -C 1.1 -D 0.05 -U 1.0 -M 0.6 -S -1.0 -c -1
Relation:     terms
Instances:    45
Attributes:   30
              Sequence, Influenza, Virus, Gene, Protein, Database, Genome, Genbank, Sequencing, Pandemic, Mutation, Genotype,
              Research, Antigenic, Pcr, Vaccine, Method, Drug, Antiviral, Resistance, Synthesis, Assay, Primer, Openflu, Array, Culture,
              Transcriptome, Proteome, Diagnostic, Phylogenetic

=== Associator model (full training set) ===

Apriori
=======

Minimum support: 0.6 (27 instances)
Minimum metric <lift>: 1.1
Number of cycles performed: 8

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9
Size of set of large itemsets L(2): 32
Size of set of large itemsets L(3): 55
Size of set of large itemsets L(4): 51
Size of set of large itemsets L(5): 27
Size of set of large itemsets L(6): 8
Size of set of large itemsets L(7): 1

Best rules found:

 1. Influenza=1 Virus=1 Protein=1 38 ==> Mutation=1 28    conf:(0.74) < lift:(1.14)> lev:(0.08) [3] conv:(1.23)
 2. Mutation=1 29 ==> Influenza=1 Virus=1 Protein=1 28    conf:(0.97) < lift:(1.14)> lev:(0.08) [3] conv:(2.26)
 3. Influenza=1 Virus=1 Protein=1 38 ==> Database=1 Mutation=1 28    conf:(0.74) < lift:(1.14)> lev:(0.08) [3] conv:(1.23)
 4. Influenza=1 Virus=1 Protein=1 Database=1 38 ==> Mutation=1 28    conf:(0.74) < lift:(1.14)> lev:(0.08) [3] conv:(1.23)
 5. Mutation=1 29 ==> Influenza=1 Virus=1 Protein=1 Database=1 28    conf:(0.97) < lift:(1.14)> lev:(0.08) [3] conv:(2.26)
 6. Database=1 Mutation=1 29 ==> Influenza=1 Virus=1 Protein=1 28    conf:(0.97) < lift:(1.14)> lev:(0.08) [3] conv:(2.26)
 7. Influenza=1 Virus=1 Protein=1 38 ==> Sequence=1 Mutation=1 27    conf:(0.71) < lift:(1.14)> lev:(0.07) [3] conv:(1.2)
 8. Sequence=1 Mutation=1 28 ==> Influenza=1 Virus=1 Protein=1 27    conf:(0.96) < lift:(1.14)> lev:(0.07) [3] conv:(2.18)
 9. Influenza=1 Virus=1 Protein=1 38 ==> Sequence=1 Database=1 Mutation=1 27    conf:(0.71) < lift:(1.14)> lev:(0.07) [3]
    conv:(1.2)
10. Influenza=1 Virus=1 Protein=1 Database=1 38 ==> Sequence=1 Mutation=1 27    conf:(0.71) < lift:(1.14)> lev:(0.07) [3]
    conv:(1.2)
```

Figure 2. Data mining results

discovered between genome and influenza. The term genome was one of the research words and was not present in any of the association rules, even though it was present on 42 of the 45 papers. Also, it was predicted that association between influenza genome and terms related directly to the application of genome databases in research of vaccine development, drug testing and development, assays and production of identification and diagnostics tests, assays for phylogenetic identification, and studies in genomics, proteomics and transcriptomics.

TABLE I. MOST INTERESTING ASSOCIATION RULES

Rules	*Confidence*	*Lift*	*Leverage*	*Conviction*
IF Mutation THEN Influenza AND Virus AND Protein	0.97	1.14	0.08	2.26
IF Mutation THEN Influenza AND Virus AND Protein AND Database	0.97	1.14	0.08	2.26
IF Database AND Mutation THEN Influenza AND Virus AND Protein	0.97	1.14	0.08	2.26
IF Sequence AND Mutation THEN Influenza AND Virus AND Protein	0.96	1.14	0.07	2.18

These unexpected results could be a consequence of the size of the corpus. However, it is not known if a corpus with all the publications from the initial results (n=76) would produce different association rules than the ones generated from the current corpus (n=45).

As in regard of the PubMed systematic research, it is possible to verify a low scientific production on the area. Previous to the search of the terms 'influenza' 'genome' 'database', a search using Medical Subject Heading (MeSH) terms was performed with fewer results, as shown in Table 2. Even with the use of generic terms instead of MeSH terms, the amount of results obtained is still considered low. This demonstrates that influenza genome data is not being used to its full potential. Furthermore, we can also question the methods that the journals and the authors are using to index their publications. Articles with an inappropriate selection of MeSH terms will result in a possible deficient recuperation of this material. The amount of information available from online influenza genome databases would provide data for research for a range of studies, from the discovery of new treatment drugs and vaccines to the widening of epidemiological studies.

TABLE II. PUBMED MESH TERMS RESEARCH RESULTS

MeSH Terms	*Number of results*
Influenza A virus AND Genome, Viral AND Databases, Genetic	26
Influenza A virus AND Genome, Viral AND Databases, Nucleic Acid	7
Influenza A virus AND Genome, Viral AND Databases as Topic	28
Influenza A virus AND Genome AND Databases, Genetic	30
Influenza A virus AND Genome AND Databases, Nucleic Acid	7
Influenza A virus AND Genome AND Databases as Topic	35
Influenza, Human AND Genome, Viral AND Databases, Genetic	8
Influenza, Human AND Genome, Viral AND Databases, Nucleic Acid	3
Influenza, Human AND Genome, Viral AND Databases as Topic	8
Influenza, Human AND Genome AND Databases, Genetic	9
Influenza, Human AND Genome AND Databases, Nucleic Acid	3
Influenza, Human AND Genome AND Databases as Topic	11

IV. CONCLUSION

In order to increase the frequency of terms and to generate different rules associating influenza genome databases and its applications, medical ontologies such as the Gene Ontology [20] and the Unified Medical Language System (UMLS) [21] can be used. Also, the use of the regular expressions (regex) system from the AntConc software can increase the efficiency of the selection of terms. Moreover, a systematic research in different electronic databases than PubMed can also increase the corpus.

The results revealed different associations than the expected and several hypotheses emerged. Mining the texts with Weka software highlighted the association between influenza virus protein and mutation sequence/database. This could be a consequence of the corpus size which is an aftermath of the insufficient amount of publications in the field.

To verify these questions, more study is necessary. More data can be obtained by expanding the search to different electronic databases and better results will be generated with the use of medical ontologies and regex. The research to assess the usage profile of influenza genome databases is important to determine its potential applications, such as studies about possible new mutated strains, and researches to develop more efficient vaccines.

REFERENCES

[1] Yang X, Yang H, Zhou G, Zhao, G. Infectious Disease in the Genomic Era. Annu Rev Genomics Hum Genet 2008;9:21-48.

[2] World Health Organization. Influenza (seasonal) fact sheet n°211. 2014.

[3] Pastore APW, Prates C, Gutierrez LLP. Implications of influenza A/H1N1 in gestational period. Implicações da influenza A/H1N1 no período gestacional. Sci Med 2012;22(1):53-8.

[4] Keitel WA, Cate TR, Couch RB. Efficacy of sequential annual vaccination with inactivated influenza virus vaccine. Am J Epidemiol 1988;127:353-64.

[5] Tripp RA, Tompkins SM. Recombinant vaccines for influenza virus. Curr Opin Investig Drugs 2008;9(8):836-45.

[6] Gao R, Cao B, Hu Y, Feng Z, Wang D, et al. Human infection with a novel avian-origin influenza A (H7N9) virus. New England J Med 2013;368:1888-97.

[7] Zhou J, Wang D, Gao R, Zhao B, Song J, et al. Biological features of novel avian influenza A (H7N9) virus. Nature 2013;499:500-3.

[8] Steinhauer DA. Influenza: pathways to human adaptation. Nature 2013;499:412-3.

[9] Ebrahimi M, Aghagolzadeh P, Shamabadi N, Tahmasebi A, Alsharifi M, et al. Understanding the undelaying mechanism of HA-subtyping in the level of physic-chemical characteristics of protein. PLoS One 2014;9(5):e96984.

[10] Jaskulski PR, Jaskulski MR, Guilhermano LG. Comparison between 1918 and 2009 flu pandemics in São Vicente de Paulo Hospital. Comparação entre as pandemias de gripe de 1918 e 2009 na perspectiva do Hospital São Vicente de Paulo em Passo Fundo, Rio Grande do Sul. Sci Med 2012;22(3):169-74.

[11] He C, Han G, Wang D, Liu W, Li G, Liu X, Ding N. Homologous recombination evidence in human and swine influenza A viruses. Virology 2008;380:12-20.

[12] Nelson MI, Simonsen L, Viboud C, Miller MA, Holmes EC. Phylogenetic analysis reveals the global migration of seasonal influenza A viruses. PLoS Pathog 2007;3:1220-8.

[13] Ananiadou S, Kell DB, Tsujii J. Text mining and its potential applications in systems biology. Trends Biotechnol 2006;24(12): 571-9.

[14] Gupta V, Lehal GS. A survey of text mining techniques and applications. J Emerg Technol Web Intell 2009;1(1):60-76.

[15] Erhardt RAA, Schneider R, Blaschke C. Status of text-mining techniques applied to biomedical text. Drug Discov Today 2006;11(7-8):315-25.

[16] Anthony L. AntFileConverter (Version 1.0.0) [Computer Software]. Tokyo, Japan: Waseda University. 2013.

[17] Anthony, L. AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. 2014.

[18] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explorations 2009;11(1):10-8.

[19] Gonçalves EC. Data mining with WEKA. Data mining com a ferramenta WEKA. III Fórum de Software Livre de Duque de Caxias. 2011.

[20] Smith B, Williams J, Schulze-Kremer S. The ontology of the Gene Ontology. Proceeding of the AMIA Symposium 2003;609-13.

[21] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32(4):281-91.