Influenza A Virus Informatics:
Genotype-Centered Database and Genotype Annotation

Guoqing Lu[1*], Kashi Buyyani[2*], Naresh Goty[2*], Ruben Donis[3#] and Zhengxin Chen[2*]

[1]*Department of Biology, University of Nebraska at Omaha, Omaha, NE 68182*
[2]*Department of Computer Science, University of Nebraska at Omaha, Omaha, NE 68182*
[3]*Influenza Division, Molecular Virology and Vaccines Branch, Centers for Disease Control and Prevention, Atlanta, GA.*

{glu3, kbuyyani, ngoty, zchen}@mail.unomaha.edu; [#] *rdonis@cdc.gov*

Abstract

Recent outbreaks of highly pathogenic avian influenza A virus infections in poultry and humans have caused considerable concerns about a future influenza pandemic in humans. In order to prepare such an unavoidable pandemic incident, effective methods for detecting and identifying dangerous virus strains that are lethal to human life must be developed. For this purpose, we developed a Web tool called FluGenome for genotyping Influenza A viruses with genome sequences. This tool can effectively detect known virus strains and identify new ones. However, it does not provide any other biological meanings to the genotypes. To annotate influenza genotypes effectively, we developed a genotype-centered database that stores various information, including sequences, genotypes, outbreak information, as well as scientific literature, and applied information retrieval and text mining techniques at the term, sentence, and abstract levels. Here we report a genotype-centered database in its design and implementation, and describe the preliminary text-mining result of influenza genotype annotation. The preliminary result demonstrated that the information retrieval and text mining techniques are valuable for the discovery of the knowledge related to influenza genotypes.

1. Introduction

Influenza is one of the most important emerging and reemerging infectious diseases, causing high morbidity and mortality in communities (epidemic) and worldwide (pandemic) [1]. The influenza virus is an RNA virus and comprises three types: A, B, and C. The type A viruses are the most virulent human pathogens among the three influenza types and cause the most severe disease. The highly pathogenic avian influenza A viruses are now causing worldwide concerns due to their recent transmission from poultry to humans, resulting in ~200 deaths from human infections [2-4].

Influenza A viruses are classified on the basis of antigenic properties of hemagglutinin (HA) and neuraminidase (NA) glycoproteins expressed on the surface of viral particles. This antigenic technique, however, has several disadvantages, including (1) the development of antisera and antigens is very time-consuming; (2) the outcome is heavily dependent on the quality of the reagents used, and (3) the assay provides qualitative rather than quantitative information [5]. Alternatively, sequence analysis, a standard technique in most laboratories, is becoming a preferred method for classification [6, 7]. A vast number of genome sequences have been generated and stored in different resources (e.g., NCBI Influenza Virus Resource) [8]. What is missing in these resources is the lack of genotype information, although such information is critical for classification and identification of new viruses [9]. This is mainly due to the complexity of the influenza A virus genome which is constituted with eight separated gene segments.

Recently, we proposed a nomenclature for naming influenza A viral genotypes [10]. This allows researchers to unequivocally describe influenza A viral genotypes to analyze, compare and communicate the molecular epidemiology of the virus. With this nomenclature, we developed a bioinformatic tool called FluGenome for influenza A viral genotyping analysis. FluGenome identifies genotypes that arose by either genetic divergence within the same host or reassortments between different circulation hosts. However, it does not provide any other biological meanings rather than "yes" or "no" answers to the predicted genotypes. Such knowledge

is largely hidden in various types of data sets, particularly in scientific literature.

As a functional description of a gene, annotation provides a useful combination of citations, comments, notations, and references that together describe all the experimental and inferred information about a gene or protein. Annotations may also be applied to the description of other biological systems. Automated, batch annotation of bulk biological sequences is one of the primary uses of bioinformatics tools. Information retrieval and text mining provide effective ways to achieve such biological annotation.

This paper continues our previous work with an emphasis of using information retrieval and text mining techniques for annotation. The remaining part of this paper is organized as follows. In Section 2 we describe the consolidated database developed by our research group. In Section 3 we describe the research methodology and in Section 4 we present the genotype annotation result. We discuss the future work of this research in Section 5, with a focus on text mining for better annotation.

2. A Genotype-centered Database for Influenza Viruses

2.1. Database design and implementation

The genotype-centered database was designed to integrate genotype data generated by our group and other existing data such as virus strains, vaccine data, outbreak data and literature information, currently scattered at several different sources into a single relational repository. Major genetic sequence databases related to influenza virus are summarized in Table 1. With respect to the genotype data, although most resources listed in Table 1 provide facilities to create phylogenetic trees, none of them attempts to store phylogenetic information, i.e., the lineages and genotypes designated for different influenza virus strains. The FluGenome database recently created by our group complements this lack [10].

Table 1. Major genetic sequence databases related to influenza A virus.

Resource	Host	Features
Influenza Virus Re-	National Center for Biotechnology Information (NCBI):	Primary flu sequence database with some basic bioinformatic
source	http://www.ncbi.nlm.nih.gov	tools.
Influenza Sequence Database (ISD)	Los Alamos National Laboratory (LANL) http://flu.lanl.gov/	Curated sequences for all users but tools for subscribed users only.
Influenza Virus Database (IVDB)	Beijing Institute of Genomics (BGI) http://influenza.genomics.org.cn	Annotated sequences with a variety of unique tools, e.g., sequence distribution and structure view.
Bio-HealthBase	http://www.biohealthbase.org/GSearch/	An integrated resource for influenza virus. Incorporating with ISD, Reactome, Immune Epitope Database, and Analysis Resource (IEDB).

To make sense of our database design, we worked with CDC virologists and devised a questionnaire and distributed it to a number of virologists. Based on the feedbacks from experts in influenza virus and our research on influenza literature, we designed a conceptual database model in such a way that facilitates not only information retrieval but also data mining and knowledge discovery. The entities (i.e., database tables) and their relationships are shown in Figure 1. The entities include Virus_Strain, Genotype, Outbreak, Vaccine, Genome, Segment, Nucleotide_sequence, Protein_Sequence, Location, Host, and Literature.
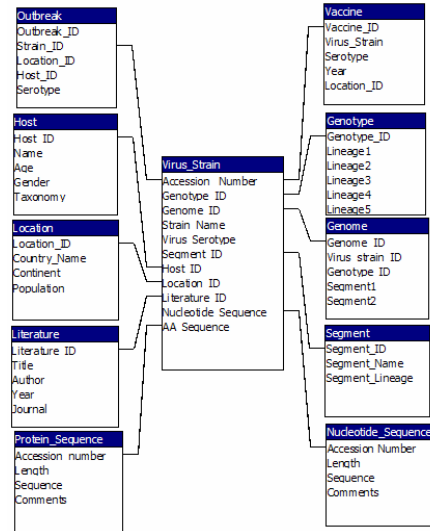

Figure 1. ER diagram shows the structure of our genotype-centered database. Primary keys highlighted and the relationships denoted with solid lines.

The database was created with the MySQL database management system, and is currently hosted on a Linux server in Dr G. Lu's lab at the University of

Nebraska at Omaha (Web address: http://bioinfo-srv1.awha.unomaha.edu) [11].

2.2. The genotype-centered database

The sequence data was downloaded from the NCBI Influenza Virus Resource (http://www.ncbi.nlm.nih.gov/genomes/FLU/), whereas vaccine data was collected from LANL ISD (http://flu.lanl.gov/). Literature data and outbreak data were collected respectively from Pub-Med database, and WHO website. Genome and genotype data were from our FluGenome database (http://www.flugenome.org). Currently there are tens of thousands of records stored in the database. Note that there is a lack of vaccine and outbreak data in the database. We are currently working on this issue.

Table 2. Number of records stored in major tables in the genotype-centered database.

Table name	# of records
Virus strain	80289
Nucleotide sequences	35839
Protein sequences	44450
Genome	2775
Genotypes	206
Vaccines	48
Outbreaks	82
Literature	1025*

*flat files, to be integrated into the database

2.3. Web interfaces to accessing genotype data and other information

Several Web pages are available for retrieving various types of information, including genome, genotype, sequence, vaccine, outbreaks, literature, database statistics, and glossary. In addition to the features described in [10], our current website supports basic queries, e.g., search by keywords, serotype, and/or country. In addition, the basic statistics of data stored in the database can be queried. In general, such statistics can be computed on the fly, but for the convenience of the use of our database, some frequently used queries have been precomputed and are readily accessed in the "Web queries" section. Figure 2 shows an example of such pre-made query.
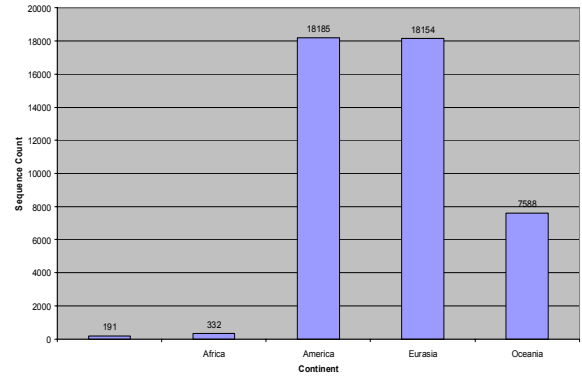

Figure 2. An example Web query shows the number of amino acid sequences available for each continent.

3. Method

We have applied information retrieval and basic text mining techniques for annotating influenza genotypes. Our method contains the collection and management of text documents as well as the annotation process itself.

3.1. Managing text documents

At the term level:

(1) Collecting documents to establish the influenza virus text base: We collected 1025 documents published in the journal Emerging Infectious Diseases (http://www.cdc.gov/ncidod/EID/) with the keywords search function in PubMed.

(2) Entity extraction: This subtask involves extraction of names of biological objects related to genotypes, such as virus, strains, etc. Porter's stemming algorithm [12] and a list of stop words are used to obtain the controlled vocabulary.

(3) Calculating TF/IDF: The key idea behind TF/IDF is that terms occurring frequently in the document (which is reflected in term frequency, or TF) but rarely in the rest of the collection (which is reflected in inverse document frequency, or IDF) are given high weights. A matrix of term TF/IDF weights has been constructed from 1,025 research articles related to influenza A virus using the similar methods as described in [13].

At the sentence and abstract levels: Sentences and abstracts provide information of interest at two dif-

ferent levels. We have developed algorithms to compute weights for each sentence (containing multiple terms) or even the entire abstract. For example, a simple way would be taking the average weight of terms appearing in one sentence or the entire abstract. Let w_{im} denote the TF/IDF weight for the mth term in document i, and k_j the number of terms appear in sentence j of this document, then the weight of the jth sentence in this document $\boldsymbol{w}_{ij}$ is calculated as

$$\boldsymbol{w}_{ij} = (\Sigma_{m=1}^{kj} \ w_{im})/k_j.$$

Similarly, if the total number of terms in document i is k, then the weight of this document $\boldsymbol{w}_i$ is calculated as

$$\boldsymbol{w}_i = (\Sigma_{m=1}^{k} \ w_{im})/k.$$

The exact formula may be revised in the future.

3.2. General operations for retrieval and analysis

Although our main objective is annotation of individual genotypes, it is important to develop several operations to provide a general picture of what these documents are about as detailed below.

(1) Retrieval of the documents: We have developed functionality for Boolean retrieval against the stored documents. Given an AND/OR query, the module constructs inverted file indices based on the keywords involved in the query and produces a ranked list as the result.

(2) Basic analysis of the documents: The following are two examples:

• Document clustering using Cluto (free software developed at University of Minnesota http://glaros.dtc.umn.edu/gkhome/views/cluto);

• Specific aspects learned from database/text base, such as using basic data mining techniques to find associations between subtypes and the host information.

3.3 Annotating genotypes

The text mining techniques have been proven of great value in the discovery of knowledge hidden in literature [14, 15]. An effective way for the biological annotation is through text mining (which is closely related to information retrieval, for storage and retrieval of unstructured data), or a combination of data mining and text mining techniques. Text mining can be used to improve the comprehensive-

ness and relevance of information retrieved from databases. Text mining can also be used to identify the elements of the infrastructure (including metadata) of a technical discipline such as influenza bioinformatics. These infrastructure elements are the authors, journals, organizations and other group or facilities that contribute to the advancement and maintenance of the discipline. Additionally, text mining can provide their specific relationships to the total technical discipline or to sub-discipline areas. Text mining can also be used to identify technical themes, their inter-relationships, their relationships with the infrastructure and technical taxonomies through computational linguistics. Based on the extracted information, it is possible to further conduct discovery from literature. There are different kinds of literature-based discovery: examining relationship between linked, overlapping literatures, and discovering relationships or promising opportunities that will not be found when read separately.

Obtaining a good annotation for genotype-centered database is a tremendous job, since it requires a wide range of machine learning and natural language processing techniques. In order to effectively handle this task, we consulted experts and conducted a thorough analysis of the overall annotation process, which results in the pipeline for annotation as shown in Figure 3.
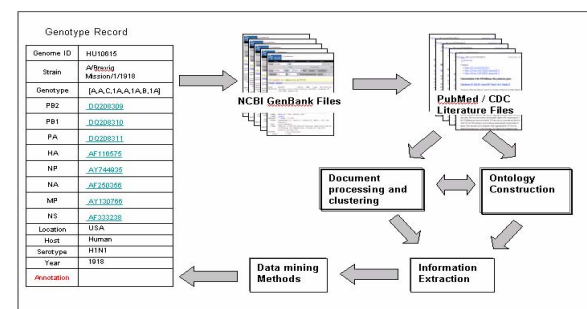

Figure 3. The pipeline of influenza genotype annotation.

The genotype entry in the FluGenome database (http://www.flugenome.org) has information on Genome ID, strain name, genotype, and GenBank accession numbers for each gene segment, host, location, serotype, and year, which provides a rich resource for text mining (Figure 3). The accession number links to the NCBI GenBank file for each genomic sequence. Within this file, there is a section called Reference, which has citations of all published journal papers describing the sequence.

We use our text mining tools to abstract reasonable meanings for each genotype. This annotation will be added to each genotype.

4. Annotation results

4.1. Documents collected for annotation

We have obtained preliminary results for a number of genotypes and will use four genotypes as an example to demonstrate results obtained at each annotation step. Information of the genotypes and associated documents is summarized in Table 3. The number of documents, i.e., journal papers, available for analyses varies from genotype to genotype, peaked at 19 for [K,G,D,5J,F,1J,F,1E]. It was revealed that this genotype is associated with avian flu, causing considerable concerns for future pandemic influenza in humans (see Table 6).

Table 3. Document information of each genotype.

Genotype	#*	PubMed ID
[A,D,B,3A,A,2A,B,1A]	8	1093066, 1538036, 1620831, 1731092, 6164798, 8774693, 9725667, 9733841
[K,G,D,5J,F,1J,F,1E]	19	12077307, 12610156, 15148370, 15235128, 15241415, 15659762, 15663858, 15681421, 15717120, 15731263, 16007072, 16306617, 16371632, 16473931, 16525739, 16532371, 16760384, 16935377, 17251574
[A,E,B,2A,A,2A,B,1A]	15	1275390, 1538036, 1643962, 1733114, 1895397, 2041090, 2398532, 2701939, 2795713, 2800339, 2974219, 520584, 6203216, 7483295, 7684877
[G,G,E,5J,F,1G,F,1E]	11	10074191, 10769072, 10920197, 11112478, 11748666, 16532371, 9430591, 9482438, 9658115, 9882316, 9927579

* number of documents.

4.2. Term-level text mining

The Venn diagram shown in Figure 4 summarizes the numbers of key terms extracted from the documents mentioned in Section 4.1 for the four example genotypes. A total of 113, 191, 166 and 163 key terms were extracted respectively for the genotypes [K,G,D,5J,F,1J,F,1E], [A,D,B,3A,A,2A,B,1A], [G,G,E,5J,F,1G,F,1E], and [A,E,B,2A,A,2A,B,1A]. On average approximately 47 terms were found in each pair of genotypes, and 19 terms found in any three genotypes. Eight terms appeared in all the four genotype. Interestingly, there are four terms "segments", "genetic", "genome", and "molecular" more biologically meaningful whereas the other four terms "analyses", "distinct", "similar", and "occurred" are relatively less meaningful. This indicates that there is still room for the improvement of our mining algorithm.
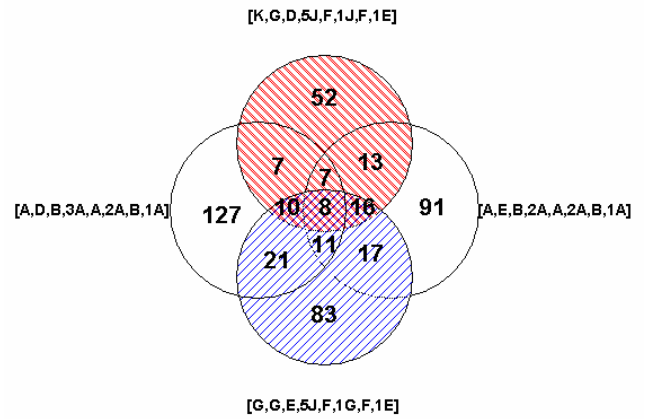

Figure 4. Venn diagram shows the number of key terms extracted from and shared between documents found to be associated with each genotype.

4.3. Sentence-level text mining

Our preliminary testing showed that a combination of two terms is more significant than other combinations, since only a few sentences were found with three or more terms combinations whereas too many sentences were found with only one term. The sentence-level mining revealed a total of 69 sentences of interest (Table 4). Interestingly, the pairwise terms employed for the mining are all biologically meaningful, which indicates these sentences are valuable for the annotation of genotypes. The weights of sentences vary from 0.09 in the paired terms "phylogenetic" and "segments" to 3.81 in "Swine" and "virus". Twenty sentences were found containing "Swine" and "virus," indicating swine is of certain significance related to influenza virus.

Table 4. Summary of sentence-level text mining.

Genotype	Terms	# sentences	Weight
[A,D,B,3A, A,2A,B,1A]	Genetic, evidence	3	0.46
	Phylogenetic, segments	1	0.09
	Nucleoprotein, polymerase	2	0.24
	Ancestral, virus	2	0.29
	Amino acid, serotypes	2	0.20
[K,G,D,5J, F,1J,F,1E]	Neurovirulence, genotypes	1	0.25
	Goose, virus	4	0.86
	Pandemic, influenza	4	1.27
	Phylogenetic, avian	3	0.41
	Respiratory, chicken	1	0.11
	Antigenic, virus	4	0.66
[A,E,B,2A, A, 2A,B,1A]	Isolate, influenza	3	0.81
	Mutations, gene	2	0.25
	neuraminidase, sequences	2	0.37
	Swine, virus	20	3.81
[G,G,E,5J, F,1G,F,1E]	Evolution, virus	6	1.01
	Glycoproteins, influenza	2	0.32
	Neuraminidase, hemagglutinin	3	0.33
	Tracheal, isolate	2	0.42
	Glycosylation, sequences	2	0.22

4.4 Abstract-level text mining

A total of 32 abstracts were found containing two or more key terms described in 4.2 (Table 5). This translates approximately 60% (32/53) of the original documents are valuable for genotype annotation. The number of terms contained varies from abstract to abstract, with a maximum value of 6. All abstracts revealed at this step are useful for the genotype annotation. However, the abstracts with more terms and with higher weights are most important; and they require more attention.

Table 5. Summary of abstract-level text mining.

Genotype	Term	PubMed ID	Weight
[A,D,B,3A, A,2A,B,1A]	Genetic, evidence, ancestral, virus	15380362	0.36
	Genetic, evidence, phylogenetic, segments, nucleoprotein, polymerase	9733841	0.29
	Nucleoprotein, polymerase	10930664	0.10
	Amino acid, serotypes, nucleoprotein, polymerase	1731092	0.12
	Amino acid, serotypes,	8774693	0.12
[K,G,D,5J, F,1J,F,1E]	Neurovirulence, genotypes, goose, virus	12077307	0.20
	Neurovirulence, genotypes	12610156	0.31
	goose, virus	16532371	0.30
	Goose, virus, antigenic, virus	16760384	0.33
	Pandemic, influenza	15148370	0.11
	Pandemic, influenza	15241415	0.18
	Pandemic, influenza, antigenic, virus	16473931	0.25
	Phylogenetic, avian, antigenic, virus	15235128	0.14
	Phylogenetic, avian, respiratory, chicken	15731263	0.10
	Phylogenetic, avian	17251574	0.10
[A,E,B,2A, A,2A,B,1A]	Isolate, influenza	12753908	0.15
	Isolate, influenza, mutations, gene	16439620	0.37
	Isolate, influenza	2701939	0.82
	Isolate, influenza, swine, virus	2800339	0.22
	Mutations, gene	2974219	0.37
	Neuraminidase, sequences	1733114	0.35
	Neuraminidase, sequences	7483295	0.62
	Swine, virus	1895397	0.18
	Swine, virus	2041090	0.09
	Swine, virus	2398532	0.07
[G,G,E,5J, F,1G,F,1E]	Evolution, virus, glycoproteins, influenza	10074191	0.19
	Evolution, virus	10769072	0.16
	Evolution, virus, neuraminidase, hemagglutinin,	9927579	0.45

	glycosylation, sequences		
	Glycoproteins, influenza, neuraminidase, hemag-glutinin, glycosylation, sequences	9882316	0.17
	Neuraminidase, hemag-glutinin	10920197	0.05
	Tracheal, isolate	9430591	0.16
	Tracheal, isolate	9482438	0.13

4.5 Making sense of genotypes

Combining automatically extracted sentences and abstracts with our manual editorial work, we obtain annotations for these genotypes. The annotation of the four example genotypes is shown in Table 6. These genotypes are associated with different subtypes and hosts. Another finding is that these genotypes arose through reassortment events and have caused either human influenza pandemics or avian influenza outbreaks (Table 6). With the text mining techniques, we are able to annotate influenza virus genotypes by assigning them the biological meanings.

Table 6. Annotation of four genotypes exampled in this report.

Genotype	Annotation
[A,D,B,3A, A,2A,B,1A]	**Subtypes:** H3N2. **Hosts:** Human and Swine. **Evolution:** Multiple reassortment events involved. **Outbreaks:** human pandemics, 1967 - 1968 Hong Kong Flu. **Representative virus strain:** A/Hong Kong/1/68. **References (PubMed ID):** 15380362, 9733841, 10930664, 1731092.
[K,G,D,5J, F,1J,F,1E]	**Subtype:** H5N1. **Hosts:** Avian, Human, Swine, Canine and Feline. **Evolution:** Reassortment events involved **Outbreaks:** Poultry in 2003-2004. Newly emerging highly pathogenic H5N1 viruses provide cause for human pandemic concern. **Representative virus strains:** A/Hong Kong/212/03 **References (PubMed ID):** 12610156, 16760384, 16760384, 15148370, 1647393, 15235128, 16473931
[A,E,B,2A, A,2A,B,1A]	**Subtype:** H2N2. **Hosts:** Human. **Evolution:** Arose through reassortment with the previous human genotype and new lineages from an unknown, but likely avian, source. **Outbreaks:** human pandemics, 1957 - 1958 Asian Flu. Representative virus strains: A/Chile/13/57. References: 2701939, 2974219, 1733114, 7483295.

[G,G,E,5J, F,1G,F,1E]	**Subtypes:** H5N1 **Hosts:** Avian and Human **Evolution:** Reassortment events involved. HA gene seems to be well adapted to domestic poultry while the rest of the genome arises from a different source. The consensus amino acid sequences of "internal" virion proteins reveal amino acids previously found in human strains, indicating these human-specific amino acids may be important factors in zoonotic transmission. **Outbreaks:** Associated with the "bird flu" incident in Hong Kong in 1997. **Representative virus strains:** A/Hong Kong/156/97. **References:** 10074191, 10769072, 9927579, 9430591, 12077307, 9482438.

5. Discussion

Although we have done substantial work, due to the complexity and the large scope of overall task, a lot of work remains to be done. There are some basic aspects related to information retrieval, such as those related to evaluation, including measurement and improvement of precision and recall.

As a particular important note, here we show evidence that the text mining techniques are valuable for biological knowledge discovery. There are a couple of issues that need to be considered for the improvement of mining results, which are: (1) how to come up with highly qualified term lists for annotation, (2) how to score the text mining results, and (3) how to abstract and produce "artificially intelligent" (AI) summary for each genotype. The first issue is essentially a sub-problem related to how to define ontology and refinement of ontologies, this is a more advanced problem of mapping/alignment of multiple ontologies for better annotation. For the second issue, currently we used pairwise terms for searching and assign a weight for each query result at the sentence level, and used two or more terms for the abstract level mining. As for the third issue, some AI techniques might be useful, which is an area worth exploring. This includes incorporation of natural language processing techniques, as well as various text mining methods such as inference of relations for genes, genotypes or other biological objects which are not in the same file.

The research reported in this paper is an ongoing long-term project. As for future directions of this research work, we would like to pursue the following tasks: (1) annotating all genotypes achieved in the FluGenome database semi-automatically (i.e., in silico mining plus manual validation); (2) developing algorithms that will reduce manual efforts in reading abstracts and sentences for the annotation of

a given genotype; and (3) disseminating the annotation results through the FluGenome website.

6. Conclusion

In this paper we described our ongoing work of developing a genotype-centered database and annotating genotypes through text mining techniques. The preliminary results demonstrated that the information retrieval and text mining techniques are valuable for the discovery of the knowledge relevant to influenza genotypes. Other related subtasks, such as development of flu ontology and document clustering, will be reported in companion papers.

Acknowledgements

References

[1] D. Morens, G. Folkers, and A. Fauci, "The challenge of emerging and re-emerging infectious diseases," *Nature*, vol. 242-249, pp. 242-249, 2004.

[2] I. Stephenson and J. Democratis, "Influenza: current threat from avian influenza," *Br Med Bull*, vol. 75-76, pp. 63-80, 2005.

[3] S. B. Muzaffar, R. C. Ydenberg, and I. L. Jones, "Avian Influenza: An Ecological and Evolutionary Perspective for Waterbird Scientists," *Waterbirds*, vol. 29, pp. 243-257, 2006.

[4] A. Moya, E. C. Holmes, and F. Gonzalez-Candelas, "The population genetics and evolutionary epidemiology of RNA viruses," *Nat Rev Microbiol*, vol. 2, pp. 279-88, 2004.

[5] R. A. Fouchier, V. Munster, A. Wallensten, T. M. Bestebroer, S. Herfst, D. Smith, G. F. Rimmelzwaan, B. Olsen, and A. D. Osterhaus, "Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls," *J Virol*, vol. 79, pp. 2814-22, 2005.

[6] H. Chen, G. Deng, Z. Li, G. Tian, Y. Li, P. Jiao, L. Zhang, Z. Liu, R. G. Webster, and K. Yu, "The evolution of H5N1 influenza viruses in ducks in southern China," *Proc Natl Acad Sci U S A*, vol. 101, pp. 10452-7, 2004.

[7] T. F. Hatchette, D. Walker, C. Johnson, A. Baker, S. P. Pryor, and R. G. Webster, "Influenza A viruses in feral Canadian ducks: extensive reassortment in nature," *J Gen Virol*, vol. 85, pp. 2327-37, 2004.

[8] S. Chang, J. Zhang, X. Liao, X. Zhu, D. Wang, J. Zhu, T. Feng, B. Zhu, G. F. Gao, J. Wang, H. Yang, J. Yu, and J. Wang, "Influenza Virus Database (IVDB): an integrated information resource and analysis platform for influenza virus research," *Nucleic Acids Res*, 2006.

[9] M. Rozanov, U. Plikat, C. Chappey, A. Kochergin, and T. Tatusova, "A web-based genotyping resource for viral sequences," *Nucleic Acids Res*, vol. 32, pp. W654-9, 2004.

[10] G. Lu, T. Rowley, R. Garten, and R. O. Donis, "FluGenome: a web tool for genotyping influenza A virus," *Nucleic Acids Res,doi:10.1093/nar/gkm365*, 2007.

[11] K. R. Buyyani, "Development of a consolidated database of influenza viruses," in *Department of Computer Science*, vol. M.Sc. Omaha: University of Nebraska at Omaha, 2006, pp. 89.

[12] C. J. van Rijsbergen, S. E. Robertson, and M. F. Porter, "New models in probabilistic information retrieval," *London: British Library*, 1980.

[13] Y. Peng, N. Yan, G. Kou, Z. Chen, and Y. Shi, "Document clustering in antimicrobial peptides research," *Proc. AMCI, 2005.*

[14] K. B. Chohen, O. Bodenreider, and L. Hirschman, *Linking biomedical information through text mining, Pacific Symposium on Biocomputing 11: 1-111*, 2006.

[15] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen, "New frontiers in biomedical text mining," *Pacific Symposium on Bocomputing*, vol. 12, pp. 205-339, 2007.