Journal of
Computer Science & Systems Biology

Chandra Sekhara Rao et al., J Comput Sci Syst Biol 2011, 4:5
DOI: 10.4172/jcsb.1000079

Research Article

Open Access

Influenza Classification from Nucleotide Sequence Database

Annavarapu Chandra Sekhara Rao* and Durvasula V. L. N. Somayajulu

Faculty of science, Department of Computer Science and Engineering, NIT Warangal, India

Abstract

In response to the wide spread of influenza H1N1, a number of alerts were issued by World Health Organization (WHO) in the form of Global Alert and Response updates [3]. Even though there is quick response in the form of alerts exists, the absence of suitable resources makes the task of the research organizations more difficult, which says there is an immediate need to provide a viable solution for reducing this gap. This also suggests Pandemic Preparedness under the Rule of International Law. A frame work is introduced to classify and construct a influenza database which is also called Influenza resource database, Data collection was especially done at Genbank, DDBJ and EMBL then Least formalized selections were Normalized and added to original collection for Classification and construction of unique local database. Classification of Influenza A was done with the RNA segmentation approach. Classification of Influenza B and Influenza C need a specific field of the database for classifying it.

Keywords: Influenza; H1N1; Classification; Nucleotide Sequence database; pre-processing

Introduction

The revised International Health Regulations for the first time in history permits an organized global response within the rule of international law and H1N1 offers the first test of its effectiveness, Even though there is quick response in the form of alerts exists, there is an enormous amount of gap exists to predict the resources required to fulfill the alerts by different research organizations [1]. Influenza is an extremely variable, fast-mutating virus. It is basically classified as Influenza virus A, Influenza virus B, and Influenza virus C. Although these three seems to be related to each other it has different disease characteristics.

In the early 21st century only we were cautioned about Influenza virus A, Interspecies transmission of influenza A viruses circulating in wild aquatic birds occasionally results in influenza outbreaks in mammals, including humans [6].

Valuable Experiments of Webby RJ, Webster RG opened the door in studying the Genetic reassortment between influenza viruses, such as H5N1and H9N2 always raises the specter of becoming dangerous for human beings if there is adaptation to man, or recombination or reassortment with human influenza viruses [5].

Health officials are reminding Americans that the H1N1 flu is still around and causing serious illness, particularly in the Southeast.Global Alert and Response system of WHO's latest Death details region wise were given in (Table 1).

Region	Deaths
WHO Regional Office for Africa (AFRO)	168
WHO Regional Office for the Americas (AMRO)	At least 8309
WHO Regional Office for the Eastern Mediterranean (EMRO)	1019
WHO Regional Office for Europe (EURO)	At least 4783
WHO Regional Office for South-East Asia (SEARO)	1769
WHO Regional Office for the Western Pacific (WPRO)	1805
Total	At least 17853

Table1 : Global Alert and Response update 97 about H1N1 by W.H.O -- As of 23rd April, 2010.

Present researchers in developing countries are facing lot of fund rising problem in the initial days of disease attack, strains of H1N1 are endemic in human, even strains responsible for the 1918 flu pandemic which killed 50-100 million people Worldwide are also important to consider this [4].

There is a need for tools and techniques for the local resource

Influenza A virus—genomic segments and coding		
RNA segment	**Designation**	**Known and probable functions**
segment1	PB2	Cap binding subunit, polymerase, virulence determinant
segment2	PB1	Catalytic subunit of RNA polymerase
segment3	PA	Subunit, viral RNA polymerase
segment4	HA or hemagglutinin	Receptor binding; membrane fusion of cell and virus to bring about infection
segment5	NP or nucleocapsid protein	Nucleoprotein (capsid) and viral synthesis
segment6	NA	Cleavescellular neuraminic acid prevents virus aggregation; facilitates release of newly produced virus
segment7	MP or M2or M1	Interacts with genome and nuclear export factor, assists viral assembly (M1).
		Tetramericion channel, controls pH in Golgi during HA synthesis and in virion uncoating (M2)
segment8	NS or NEP or NS1	Post-transcription RNAcontrol;interferon antagonist(NS1)
		Nuclear export of viral RNA, viral assembly(NEP)

Table 2: Influenza virus data classifications with RNA segmentation approach.

***Corresponding author:** Annavarapu Chandra Sekhara Rao, Faculty of science, Department of Computer science and Engineering, NIT Warangal, India, E-mail: chandra_cse@nitw.ac.in

Received April 27, 2011; **Accepted** November 22, 2011; **Published** November 26, 2011

Citation: Chandra Sekhara Rao A, Somayajulu DVLN (2011) Influenza Classification from Nucleotide Sequence Database. J Comput Sci Syst Biol 4: 077-080. doi:10.4172/jcsb.1000079

preparations at every research institute in order to overcome delays in the research database receiving and also need for up-to-date information at research centers, Our approach provides the capability to identify standard Data classification schemes that segments the occurrences of Influenza A virus based on the RNA segment type, Other possible parameter based classification for Influenza B and Influenza C.

Influenza virus known and purported functions of the virus-encoded proteins are listed in (Table 2) [2].

Frame work

A frame work is designed and developed for classifying and constructing a local database with various Online Nucleotide Sequence databases like GenBank, DDBJ and EMBL from least formalized selections. We can see proposed frame work in (Figure 1). Frame work begins by giving Influenza search query for the GenBank ,DDBJ and EMBL databases which are our proposed databases for testing our frame work, then retrieving a selection of Nucleotide Sequence that correspond to the query. We can see this in detail in material and method section.

As online results are in list and XML formats in their own styles, solving this problem need some attention in getting the data and preprocessing it. We can see these preprocessing algorithms in material and method section in detail.

Preprocessing should be done separately to proposed databases. After the tables have been formed separately to proposed databases with preprocessing step, then grouping of all the records data from three should be taken to single Excel file which is like shown in (Figure 3). This step is the major step sometimes we can treat this as a resource database.

Once the local database is available in order to deal easily with available data we need to further divide them in to small related collection which here we are taking segment wise collection. We can also use any key parameter of the database for this type of classification.

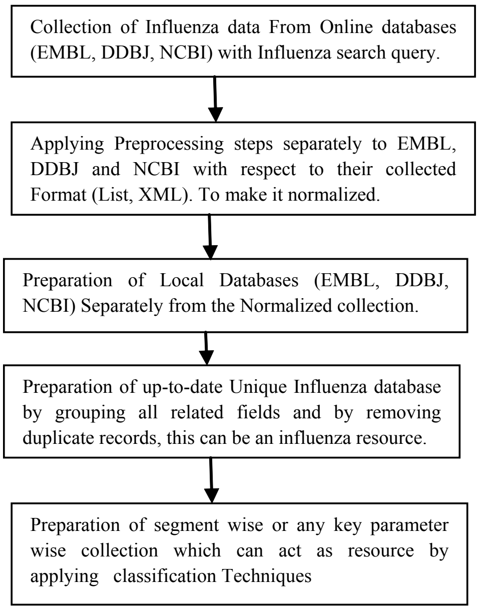
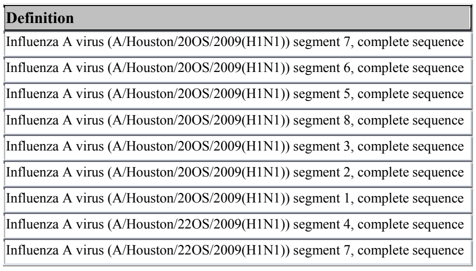
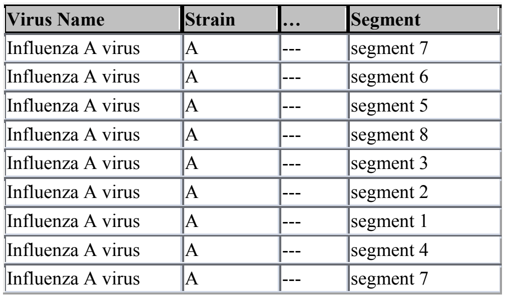

Collection of Influenza data From Online databases (EMBL, DDBJ, NCBI) with Influenza search query.
Applying Preprocessing steps separately to EMBL, DDBJ and NCBI with respect to their collected Format (List, XML). To make it normalized.
Preparation of Local Databases (EMBL, DDBJ, NCBI) Separately from the Normalized collection.
Preparation of up-to-date Unique Influenza database by grouping all related fields and by removing duplicate records, this can be an influenza resource.
Preparation of segment wise or any key parameter wise collection which can act as resource by applying classification Techniques

Figure 1: Frame work of the designed work and flow of important actions.



Definition
Influenza A virus (A/Houston/20OS/2009(H1N1)) segment 7, complete sequence
Influenza A virus (A/Houston/20OS/2009(H1N1)) segment 6, complete sequence
Influenza A virus (A/Houston/20OS/2009(H1N1)) segment 5, complete sequence
Influenza A virus (A/Houston/20OS/2009(H1N1)) segment 8, complete sequence
Influenza A virus (A/Houston/20OS/2009(H1N1)) segment 3, complete sequence
Influenza A virus (A/Houston/20OS/2009(H1N1)) segment 2, complete sequence
Influenza A virus (A/Houston/20OS/2009(H1N1)) segment 1, complete sequence
Influenza A virus (A/Houston/22OS/2009(H1N1)) segment 4, complete sequence
Influenza A virus (A/Houston/22OS/2009(H1N1)) segment 7, complete sequence

Figure 2: Data before splitting of Definition field.



Virus Name	Strain	...	Segment
Influenza A virus	A	---	segment 7
Influenza A virus	A	---	segment 6
Influenza A virus	A	---	segment 5
Influenza A virus	A	---	segment 8
Influenza A virus	A	---	segment 3
Influenza A virus	A	---	segment 2
Influenza A virus	A	---	segment 1
Influenza A virus	A	---	segment 4
Influenza A virus	A	---	segment 7

Figure 3: Data after splitting of Definition field.

Materials and Methods

Designed Framework procedure begins by giving Influenza search query for the proposed databases and retrieving a selection of Nucleotide Sequence's that correspond to the query. As the three Nucleotide Sequence databases were different in their representation and output styles we need to gather information from the search query with different procedural steps.

From proposed database, obtained Results were processed via an exchange buffer into the table of the Excel file, with the designed macros lists were processed in sequential order and place data into corresponding cells. Each Nucleotide Sequence in the list is of the proposed database differently associated with differently set of rows.

In EMBL we can observe the following annotations in the List format.

- Access numbers for the given sequence.
- Definition of that Nucleotide Sequence.
- Viewing style web links.
- Reference site links

In Genbank we can observe the following annotations in the List format

- Definition of that Nucleotide Sequence.
- Sequence length, topology, and molecular type.
- Version number, alternative access numbers for the given sequence in other databases.

Note: GenBank have a feature of storing query results as XML file.

In DDBJ we can observe the following annotations in the List format. By default Each Nucleotide Sequence in the list is represented with single row consisting of three fields, additionally there is facility to show 3 more fields information. Default annotations are as follows.

- Access numbers for the given sequence.
- Definition of that Nucleotide Sequence.
- Sequence length.

Note: DDBJ have a feature of downloading the XML results through a link via Email which is provided when search query is given.

As online results are in list and XML formats in their own styles, solving this problem need some attention in getting the data and preprocessing it.

In List format main problem was that the first row of Genbank, second row of EMBL and second field of DDBJ in the annotated sequence list contained information in the least formalized form and did not readily lend itself to any universal processing algorithm. In XML main problem was that the definition field of sequence table in the least formalized form and did not readily lend itself to any universal processing algorithm. We have different approaches to process List format and XML format separately to provide resource for the search query.

With List format Approach processing of lists into several steps with each step executed by VBA macros separately to proposed databases. A separate list of terms is formed at each step separately to EMBL, DDBJ and GenBank. The list is usually comprised of words that are closely related, e.g. that represents names of species, strains, or genes.

After the tables have been formed separately to proposed databases, then grouping of all the records data from three Databases should be taken to single Excel file. This list enables the user to determine the number of related records for each gene and each corresponding species which can be treated as Local useful collection. We convert Excel data in to Access Database then remove duplicate records based on the Accession number, after removal of duplicate data we then categorize sequences into groups, Easiest is processing with XML, we can create Access database for the resultant XML output. Remove & edit the fields and tables in the database according to necessity to create local database.

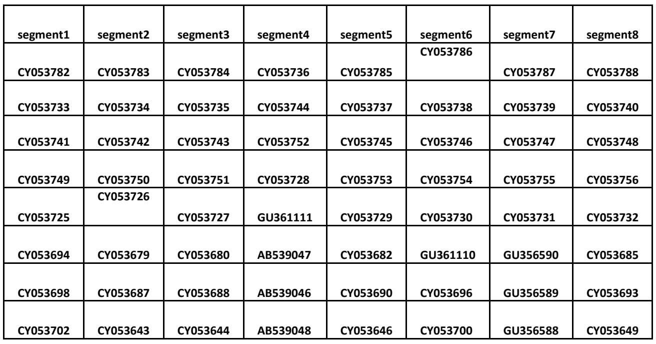
Experimental Results

Once the retrieved selection of Nucleotide Sequence's were collected, with the preprocessing steps convert the selections to intermediate stage of results as shown in (Figure 2), then identify fields that are in least formalized format split or process them in required format, this process was done separately to list and XML formats, club them to original source of collection, we used H1N1 data for implementation of proposed approach.

Process of splitting

The main problem is splitting of least formalized data and making them formalized, we have to identify the fields which are in least formalized form, then with VBA macro or UNIX scripting split or process them in to sub fields then add this divided or processed data to the original. In (Figure 2) we have H1N1 Nucleotide Sequence collections from proposed databases; definition field was identified as least formalized field. At the end of the splitting process we can see the least formalized selections as a formalized collections.

Note: all this process should be done separately to GenBank, DDBJ and EMBL



segment1	segment2	segment3	segment4	segment5	segment6	segment7	segment8
CY053782	CY053783	CY053784	CY053736	CY053785	CY053786	CY053787	CY053788
CY053733	CY053734	CY053735	CY053744	CY053737	CY053738	CY053739	CY053740
CY053741	CY053742	CY053743	CY053752	CY053745	CY053746	CY053747	CY053748
CY053749	CY053750	CY053751	CY053728	CY053753	CY053754	CY053755	CY053756
CY053725	CY053726	CY053727	GU361111	CY053729	CY053730	CY053731	CY053732
CY053694	CY053679	CY053680	AB539047	CY053682	GU361110	GU356590	CY053685
CY053698	CY053687	CY053688	AB539046	CY053690	CY053696	GU356589	CY053693
CY053702	CY053643	CY053644	AB539048	CY053646	CY053700	GU356588	CY053649

Figure 4: Segment wise collections of processed nucleotides sequences of H1N1.

Preparation of unique database

After collecting required data separately from Genbank, DDBJ and EMBL, Immediate step is to prepare a unique source of Nucleotide Sequence database for all the collection and remove any duplicate records. This database can be treated as research database.

Classification and resource database

Further division of the local database is also good if we know it is important. Classification is one of the approaches for doing this. Classification of Influenza A is with RNA segmentation approach Classification of Influenza B and Influenza C were depends on the interest of bioinformatics researcher, there is possible for species wise collection or sub species wise collection or Date wise collection or any key feature wise collection from the processed collection. Here in (Figure 4) we showed segment wise collections of processed nucleotides sequences as Influenza resource.

Conclusion and Future Work

Instead of Having one Nucleotide Sequence Database Collection, Up-to-date collection and classification of influenza virus of All Nucleotide Sequence Database Collection is possible, esp. RNA segment wise collection in Influenza A virus, species wise or any defined parameter wise collection or classification in Influenza B and Influenza C. Prepared collections can be treated as Local Database which is very much useful for bioinformatic analyst to apply his ideas to get the key information in the research; we can also treat it as Flu Research database.

In future there is a scope for Structural Characterization on influenza, Classification scope in influenza B and Influenza C with respect to purported definitions available. Even we have plenty of flu resources and any other useful literature we still need lot of bioinformatic analysis for the better future society. As Nucleotide Sequence databases grow rapidly in volume, there is a chance for growth in influenza also, if huge influenza local database available there may be need for data mining techniques for processing it. We also can extend this frame work to other online protein sequence databases or any online bioinformatic databases.

References

1. Gostin LO (2009) Influenza A (H1N1) and Pandemic Preparedness Under the Rule of International Law. JAMA 301: 2376-2378.
2. Hilleman MR (2002) Realities and enigmas of human viral influenza: pathogenesis, epidemiology and control. Vaccine 20: 3068-3087.

Citation: Chandra Sekhara Rao A, Somayajulu DVLN (2011) Influenza Classification from Nucleotide Sequence Database. J Comput Sci Syst Biol 4: 077-080. doi:10.4172/jcsb.1000079

3. Global Alert Response of WHO latest Updates.

4. Palese P (2004) Influenza: old and new threats. Nat Med 10: S82-87.

5. Webby RJ, Webster RG (2001) Emergence of influenza A viruses. Philos Trans R Soc Lond B Biol Sci 356: 1817-1828.

6. Matrosovich M, Tuzikov A, Bovin N, Gambaryan A, Klimov A et al. (2000) Early alterations of the receptor-binding properties of H1, H2, and H3 avian influenza virus hemagglutinins after their introduction into mammals. J Virol 74: 8502–8512.

7. Jiawei Han , Micheline Kamber Data Mining: Concepts and Techniques. 2nd Ed.

8. Advances in Data Mining Applications in Image Mining, Medicine and Biotechnology, Management and Environmental Control, and Telecommunications, eBook ISBN: 3-540-30185-2

9. Knowledge Discovery and Emergent Complexity in Bioinformatics, Lecture Notes in Bioinformatics 4366. Springer- Verlag Berlin Heidelberg 2007.

10. Mohammed J. Zaki (2005) Data Mining in Bioinformatics (Advanced information and knowledge processing). Springer- Verlag London Limited.