ISSN 2047-3338

Toward the Influenza Virus Data Mining from a DNA Data Bank

Nguyen Gia Khoa[1], Tran Van Lang[1], Tran Van Hoai[2], Van Dinh Vy Phuong[3]

[1]Institute of Applied Mechanics and Informatics, Vietnam Academy of Science and Technology, Vietnam
[2]HCMC University of Technology, Vietnam National University in HCM City, Vietnam
[3]Lac Hong University, Vietnam Ministry of Education and Training, Vietnam
nguyengiakhoa@yahoo.com, tvlang@vast-hcm.ac.vn, hoai@cse.hcmut.edu.vn, phuong@lhu.edu.vn

***Abstract*—Influenza virus is one of the causes of flu in human beings as well as in animals. Vietnam, with its tropical climate, is one of the countries located in heavily affected areas of the influenza virus. Building a localized information system of flu viruses for Vietnam provinces is quite necessary not only for research but also for other communities. The paper presents such an information system to extract the local viruses from international databases and provides certain utilities for research and non-research communities. The first of the paper presents an overview of data banks of flu virus in the world. The second part shows the methods of building the information system on the influenza virus. The final parts will present some results obtained and conclusion.**

***Index Terms*—Bioinformatics, Networking and Knowledge Discovery**

I. INTRODUCTION

INFLUENZA is the reason causes to flu sickness at human and animals. Influenza was classified into 3 different types including flu A, flu B and flu C. In which, flu A contents various subtypes (H1N1, H5N1, etc.), which is the most popular and dangerous kind of influenza. The virus is easily transmitted from animal to animal, from animal to person, and especially from person to person as well, it is one of the most dangerous kinds of virus, affected to the economy as well as human's health ever before.

Presently, a huge molecular weight of influenza has been encrypted and filled in general data foundation globally such as NCBI (National Center for Biotechnology Information) [2]. According to statistic, NCBI was storing more than 150,000 sequence of influenza, which was collected and encrypted from many countries all over the world during last period.

In Asia, Beijing Institute of Genomics, China was contributed a data of IVDB (Influenza Virus Database http://influenza.psych.ac.cn/). IVDB stored about 43,000 sequences of many kinds; different types of influenza of many other countries.

However, the information provided by NCBI, IVDB systems is only got details at national level. It means that there has no provincial level of information in one country.

Vietnam got tropical climate, has been one of the worst - affected countries of influenza. Since 2003, Vietnam had to face with H5N1 disease, also known as avian flu, which was caused dead in humans and a huge amount of animal species, had been destroyed with total hundreds and millions of USD's loss.

Due to the high – important danger of influenza, a series of analysis had been done in Vietnam. The strong development of biologic technology helped Vietnam could process analysis at the level of molecular biology. The encryption of sequences or even the genomes of influenza had been done during last period. At present, there has over 2,951 sequences of influenza were encrypted in many provinces since 2001.

Department of animal health processed many researches about influenza, especially avian influenza H5N1, in which the research's group of Dr. Nguyen Tien Dung was decoded total genomes of 33 virus at many different provinces from 10/2005 to 05/2007 such as: Dong Thap, Soc Trang, An Giang, Ha Tay, Vinh Long, Ha Noi. The research's group has pointed out the relationship among avian influenza H5N1 of many different provinces in Vietnam [4].

Another research's group of Dr. Le Sy Vinh at Hanoi National University - University of Engineering and Technology, they have developed methods and bioinformaticstoolsin order to analyze influenza data. The group has exposed a model altering amino acid of influenza, which could enhance the accurate when analyzing protein's sequence in comparison with previous models.

Our research group, leader by Assoc. Prof. Dr. Tran Van Lang at Institute of Applied Mechanics and Informatics, has also passed many years for analyzing and contributing the bioinformatics tools that severed for researching sequence, a foundation of analyzing bacterium and virus. Some typical software such as multiple sequence alignment, drawing plasmid map, designing enzyme cutting imitation software, taxonomy building [5], [6].

Despite there have many researches of influenza in Vietnam, however, those researches are mainly focused on encrypting DNA sequence and protein, so that they could analyze the relationship among them.

A problem was raised that: A research will be constructed to automatically gather data from database of gens, relating to influenza from biologic data bank globally. From that point, an updated data system of influenza will be constructed for provinces of Vietnam. This data system could provide information for researchers, managers (ministry, public health industry); therefore people could get information on influenza data at many provinces of Vietnam.

There are 4 sections in this paper, in the first to describe the overview of data banks of flu virus in the world. The second section shows the methods of building the information system on the influenza virus. The experimental results were presented in third section, and the final part will present some conclusions and future works.

II. BUILDING AUTOMATIC UPDATING DATA SYSTEM OF INFLUENZA

A. *Challenges when automatic updates*

As usual, users will use finding tool and get systems sequence of NCBI or DDBJ (DNA Database Bank of Japan) when they need any virus data. Users could find and get out sequences when using keywords through these systems. Then, users will get results including a list of sequences, and they will choose essential ones and permitted by data system. In case of many results (thousands) for one finding word, users must pay a lot of time in order to make choice with needed sequences. However, numbers of sequences were continuously encrypted, so that it would take long time and manipulation for getting new sequences. The question is that how users could get automatically sequences. That means they must construct a system, which could discover new sequences and automatic get them out.

B. *Automatic updating data system model*

For creating conveniences when manages, analyses and automatic updates database, there need to be timely and regularly perform in order to support for updating process to get easy, high effective and lasted virus database on over the world.

To achieve those criterions, the model permits automatic updating influenza system has build (Fig. 1)

Remarks: Banks of huge database systems on the world such as NCBI, DDBJ, EMBL (Nucleotide Sequence Database - http://www.ebi.ac.uk/embl/)often be used and proclaimed researched data by biologists. Database exchanges among NCBI, DDBJ and EMBL were formed International Nucleotide Sequence Database Collaboration - INSDC. Consequently, this database always contains new information.

C. *Description system operation*

These are two main systems that automatically update database of influenza and connect for taking out data. NCBI supply files that could access to general information about nucleotide sequence, protein of influenza, which were encrypted on the world. Information of this system is always updated [2]. When a biologic sequence is published, INSDC banks will allocate for that sequence a unique access digital code, called ACESSION of sequence. With this information, users could access details of sequence content that were analyzed and published by laboratories through accessing code.

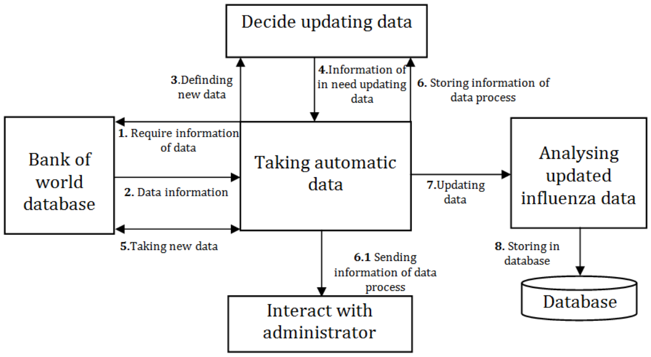

Decide updating data
3.Defînding new data
4.Information of in need updating data
6. Storing information of data process
Bank of world database
1. Require information of data
2. Data information
5.Taking new data
Taking automatic data
7.Updating data
Analysing updated influenza data
8. Storing in database
Database
6.1 Sending information of data process
Interact with administrator

Fig. 1. Model of automatic updating data system

Information system of influenza from NCBI delivers fully general information of influenza sequence, which was encrypted on over the world.

Access address: ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/, supply following files:

- genomeset.dat – File contains general data of gen
- influenza_na.dat – File contains general data of nucleotide
- influenza_aa.dat – File contains general data of protein

DDBJ has information system sufficiently serves for bioinformatics researchers. One of usefully systems is WABI (Web API for biology) [9],[10]. Through API function of this system, users could realize quantity and ratio of A, T, G, C of a biologic sequence and get detailed content of a sequence from host server of DDBJ. For example, taking a biologic sequence with any accesstion code, we could do:

Prototype of DDBJ:

http://xml.nig.ac.jp/rest/Invoke?service=GetEntry&method=getDDBJEntry&accession=<ACCESSION>, in which ACCESSION is the code of needed sequence.

With ACCESSION is GU186747, we have following inquiry function:

http://xml.nig.ac.jp/rest/Invoke?service=GetEntry&method=getDDBJEntry&accession=GU186747

With below results:

LOCUS*GU186747* 1372 bp cRNA linear VRL 25-NOV-2009

DEFINITION Influenza A virus(A/Muscovy duck/Ca Mau/07-04/2007(H5N1))segment 6 neuraminidase(NA)gene,complete cds...

Remarks: With an accession code of sequence, combine with API functions of DDBJ, users could take total content of sequence out.

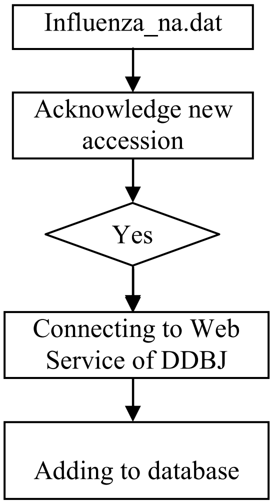


Fig. 2. Process of getting new data

As description above, automatic updating system of influenza will rely on survey information about influenza data on NCBI system. Base on survey information in file influenza_na.dat, system will identify necessary data. After that, system will record accession code of sequences. Following, system will active connected facility with DDBJ system in order to get sequences and update in intrinsic database.

Fig. 2 will indicate clearly about getting new data process.

III. EXPERIMENTAL RESULTS

A. *Overall information*

Influenza information system was constructed into 2 components. In which first component is a background application assumes the updating database of influenza from biologic database bank on the world. The second component is a web application, undertakes supplying influenza information to community of researchers, managers and people.

When making experiment, system has collected A, B, C sequences fro DDBJ and NCBI. Initially, the system has collected more than 140,000 influenza sequences from 130 countries. Especially, the system had assembled fully data of influenza of Vietnam, which was declared on public biologic data bank on the world with the amount of 2,951 sequences.

It was named IVDBVN (Influenza Virus Database of Vietnam)

B. *Achievement results*

Supplying influenza information with many criterions:

For serving researched activities as well as learning about information of influenza sequences, IVDBVN enables users studying and collecting as many criterions. Users only choose parameters, and then results will be displayed immediately. IVDBVN also permits users taking sequences according to formatting of NCBI system such as Fasta, Accession List, and ProteinList. For example: Getting information of influenza sequences on the map point – Tien Giang Province, IVDBVN will give following information:

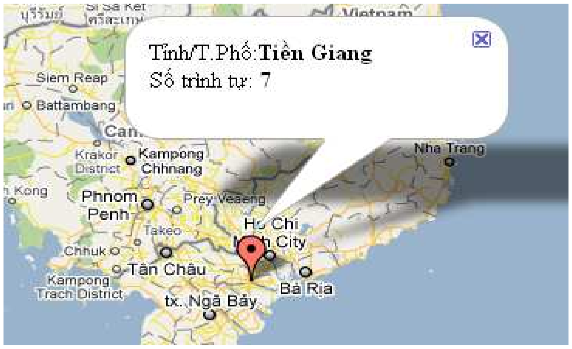
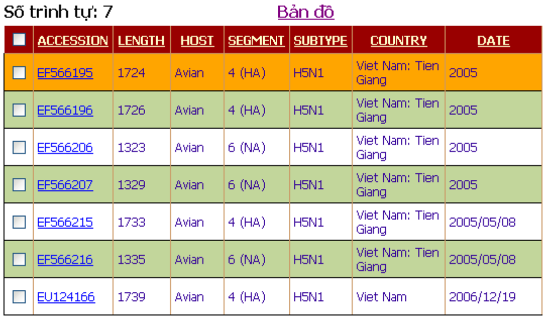


Fig. 5. Influenza sequences of Tien Giang Province's neighbors



Số trình tự: 7 Bản đồ

□	ACCESSION	LENGTH	HOST	SEGMENT	SUBTYPE	COUNTRY	DATE
□	EF566195	1724	Avian	4 (HA)	H5N1	Viet Nam: Tien Giang	2005
□	EF566196	1726	Avian	4 (HA)	H5N1	Viet Nam: Tien Giang	2005
□	EF566206	1323	Avian	6 (NA)	H5N1	Viet Nam: Tien Giang	2005
□	EF566207	1329	Avian	6 (NA)	H5N1	Viet Nam: Tien Giang	2005
□	EF566215	1733	Avian	4 (HA)	H5N1	Viet Nam: Tien Giang	2005/05/08
□	EF566216	1335	Avian	6 (NA)	H5N1	Viet Nam: Tien Giang	2005/05/08
□	EU124166	1739	Avian	4 (HA)	H5N1	Viet Nam	2006/12/19

Fig. 6. Displaying influenza information of Tien Giang Province's neighbors

Displaying influenza information of Vietnam on Google Map:

In accordance of Google map appearance; this technology has been applied into displaying general information for influenza at provinces of Vietnam on the map. With initial outcomes, our system has performed on 2 forms of map: to mark and make relationship between color, quantity of sequences (Fig. 7). These facilities are only appeared on IVDBVN system. Especially, with the map forms displays according to relationship between color and quantity will make visual information for people who get influenza information of neighborhood provinces. Thereby, the system will help people get certain understanding about the distribution as well as propagation of influenza.

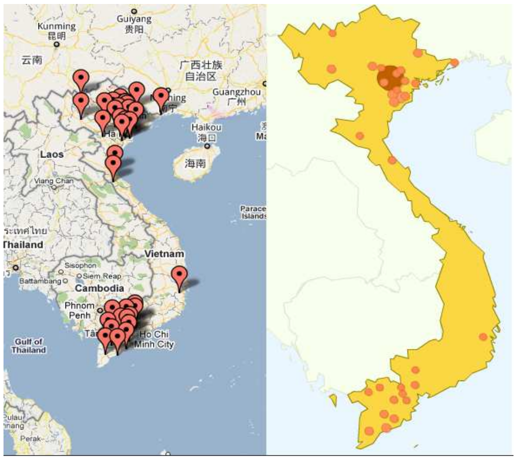


Fig. 7. Displaying influenza virus of Vietnam on Google map

Influenza's statistic of Vietnam:

(Statistic according to sequences' quantity allocation of provinces)

This facility is only appeared on IVDBVN system. Through this graph, users could see sequences' quantity of influenza at provinces of Vietnam (Fig. 8).

Sub-group of influenza follow provinces is classified into attribution /country in the content of sequence file. For example: /country="Viet Nam: Tien Giang" will get sequences of influenza at Tien Giang, Vietnam.

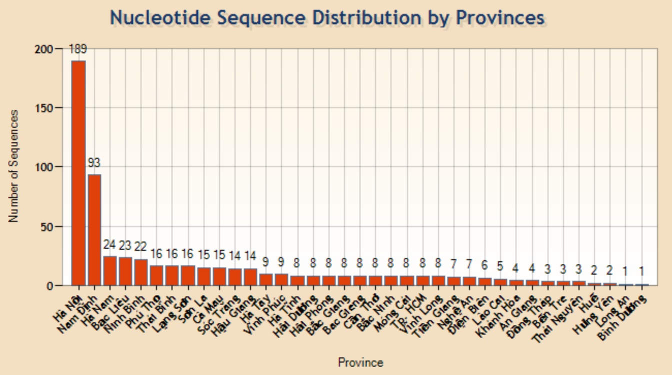


Fig. 8. Nucleotide sequence distribution by provinces

Statistic quantity according to sequences' subtypes:

Main subtype of influenza in Vietnam is H5N1 (Fig. 9). In comparison with other subtypes on the world, there is different, while the popular subtype is H1N1. In addition, IVDBVN system also permits users making statistic of influenza allocation follow subtypes, gene segments, hosts, time, dead cases on total infected cases for countries.

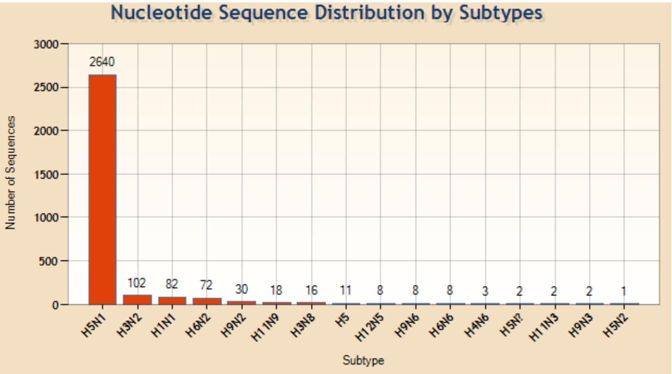


Fig. 9. Nucleotide sequence distribution by subtypes

Statistic sequences' quantity according to segments:

The popular gene segment of influenza on the world is HA, seen by this graph (Fig. 10). This is also the most general gene type of influenza on the world.

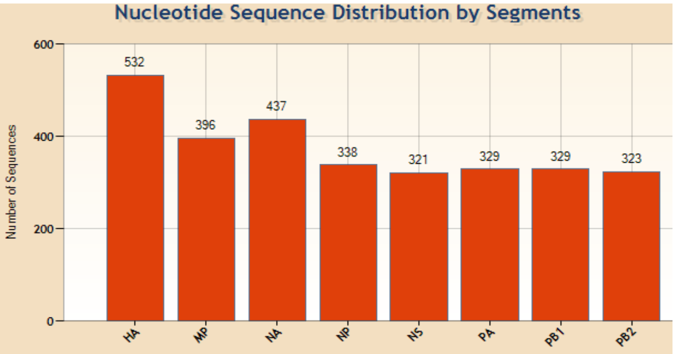


Fig. 10. Nucleotide sequence distribution by segments

Statistic quantity sequences according to hosts:

A kind of hosts – Civet is only existed in Vietnam. If host is environment, which means influenza will be found in the atmosphere environment of Vietnam. This fact gives the result that the spreading influenza propagation in Vietnam is rather high.

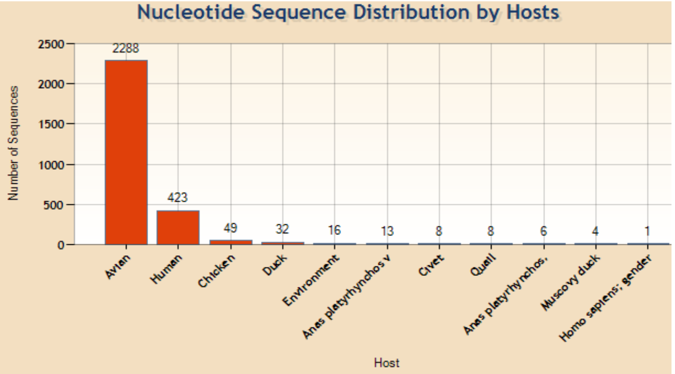


Fig 11. Nucleotide sequence distribution by hosts

Statistic collected and public quantity sequences yearly:

This facility could help users collect quantity of public sequences as critical time, especially in 2008, 2009, 2010 the numbers of proclaim sequences are extremely high. Simultaneously, 2009 is also the year that World Health Organization (WHO) announced flu pandemic A/H1N1 (Fig. 12).

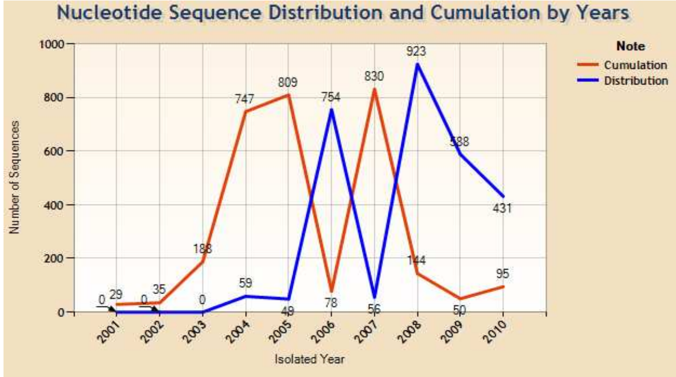


Fig. 12. Nucleotide sequence distribution and accumulation by years

Statistic dead case on total infected victims A/H5N1:

This is a useful facility only on IVDBVN system. The graph will help users get information about dangerous level and flu pandemic at countries as critical time.

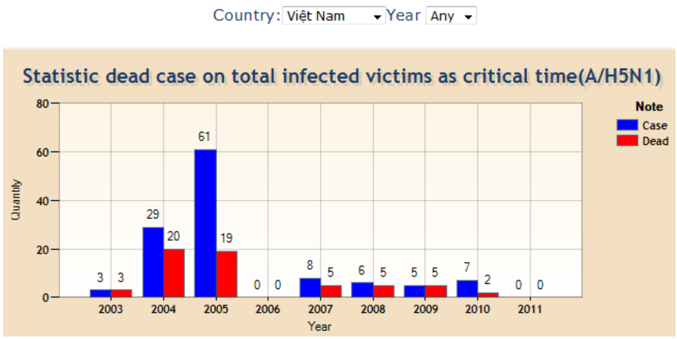


Fig. 13. Statistic dead case on total infected victims as critical time

Data Mining of influenza:

Introduction

Flu pandemic situation at every country has characteristic nature. Influenza sequences will be filled this specific information. When analyzing the association among item in sequences of influenza database, productive information will be got. For instance, in Vietnam "74% of influenza sequences in Vietnam are belong to H5N1 subtype and host is avian"; "over 92% of influenza sequences collected in December to know by H5N1 subtype and host is avian". Or another example in Indonesia "over 98% of sequences collected in January are belong to H5N1 and host is human".Thus, virus subtypes will cause flu pandemic at every country, host of virus, location, time are different.

The extraction of hidden information from database of influenza sequences will help researchers and managers get useful information. Base on this information, they could make out forecasts as well as preparation for the flu pandemic.

Data Mining of influenza in Vietnam:

In order to extract useful information of influenza in Vietnam, this paper uses the association rule in data mining. There are many methods for finding frequent itemsets and this article employs IT-tree method (Itemset Tidset – Tree). This method attains high efficiency than Apriori methods [7], [8]. On the process of finding popular files, IT-Tree only scans database for one time and discover minimum supporting thresholdfrequent itemsets. Moreover, creating frequent itemsets by IT-Tree method is not arising candidate itemsets. These specific natures are very important when developing facilities on Web application.

Data used in mining purpose is subjected to influenza data of Vietnam. With the target of finding general rule, which help specific researchers and managers, acknowledge the linkage among subtypes, hosts, provinces and pandemic time. Base on these relationship, they could active in respond with pandemic situation. Parameter for the problem: Minsupport = 10%; Mincofidence = 50%, following results:

Table 1. List of influenza association rule at provinces in Vietnam

Left side	Right side	Support	Confidence	Meanings
H5N1	Avian	17%	100%	100% influenza with H5N1 subtype has host as Avian.
H1N1,Hanoi	Human	33%	100%	100% influenza with H1N1 subtype in Hanoi has host as human.
January, Hanoi	Human	11%	93%	93% influenza collected in January at Hanoi has host as human.
January, Namdinh	Avian	16%	100%	100% influenza collected in January at Namdinh has host as Avian.
Human,H3N2	Hanoi	16%	98%	98% influenza with H3N2 subtype has host as human, origin in Hanoi.
Namdinh	Avian	20%	100%	100% influenza in Namdinh has host as avian.

Table 2. List of influenza association rule in Vietnam

Left side	Right side	Support	Confidence	Meanings
Human	H5N1	22%	52%	52% influenza has host as Human subjected to H5N1 subtype
Avian	H5N1	30%	62%	62% influenza has host as Avian subjected to H5N1 subtype
H3N2	Human	12%	100%	100% influenza with H3N2 subtype has host as human.
H6N2	Avian	11%	100%	100% influenza with H6N2 subtype has host as Avian.
October	Avian	11%	69%	69% influenza collected in October has host as Avian.
December	H5N1	30%	93%	93% influenza collected in December subjected to H5N1.
December	Avian, H5N1	24%	74%	74% influenza collected in December has host as Avian and subtype as H5N1
December,Avian	H5N1	24%	100%	100% influenza collected in December has host as avian subjected to H5N1 subtype.
H1N1	Human	12%	100%	100% influenza with H1N1 subtype has host as Human.

Remarks: Through Table 2, the following useful information will be derived:

- If flu pandemic happens at the end of the year (October, December), hosts are usually human or avian.
- If flu pandemic happens in Hanoi, Namdinh in January, hosts are human or avian.
- Due to the parameter of Minsupp is 10%; provinces of Vietnam have rarely appeared at available situations.

C. *Comparison and evaluation*

Comparison with influenza information system (Table 3).

Table 3. Comparison and Evaluation with influenza information system

Ord	Function	Systems		
		NCBI	IVDB	**IVDBVN**
1	Storing sequences	>150,000	>43,000	>140,000
2	Storing influenza sequences in Vietnam (*)	2,951	1,077	2,951
3	Providing influenza information at National Level	Yes	Yes	Yes
4	Providing influenza information of neighbor countries	No	No	Yes
5	Providing influenza information at provincial level	No	No	Yes
6	Solving bioinformatics mathematic problem	Yes	Yes	No
7	Displaying information on the map	No	Yes	Yes
8	Statistic as graph	No	Yes (**)	Yes
9	Taking sequences out	Yes	Yes	Yes
10	Statistic dead case on total infected victims	No	No	Yes
11	Data mining of influenza	No	No	Yes

(*) Statistic as of 22/03/2011; (**) no parameters

Remarks: Through the above table, IVDBVN has provided a number of useful facilities for users such as:

- Providing influenza information to provincial level of country. At present, the system has displayed information of 32 provinces in Vietnam.
- Displaying the distribution of influenza under graphic shape by utilized Google map. The map could display the differences between the numerous and few infected regions. Thereby, people could certainly acknowledge of the influenza distribution.
- Providing influenza information of neighbour provinces, this facility could help researchers and managers could make comparison as well as evaluation about the relationship of influenza subtypes.
- Providing influenza information at the form of graph, IVDBVN system permits users setting up parameters for graph. So that, users could easily get visual information with different criterions.
- The facility of data mining of influenza, the extraction of hidden information sequences of countries will help researchers and managers get useful information. With these data, they could active give out forecasts as well as preparation in respond with flu pandemic.

IV. CONCLUSION AND FUTURE WORKS

The extremely development of biologic technology in Vietnam has helped us with many researches of influenza at molecule biology level. The encryptions of sequences or even genomes of influenza have been proceeding during last period.

As usual, sequences that filled at international database such as NCBI, DDBJ to share with users. However, most of data filled at those systems contained with general information, not having details at provincial level. Therefore, we have insufficient information about the spreading of influenza and particularly analyzing in Vietnam. So that, the building an automatic system which could update influenza data from data bank of biology on the world (Fig. 1); providing particular data at provincial level is necessary.

Providing statistical data tools are very essential. This function will help us understand about the distribution of influenza on every country, provinces of Vietnam, distribute by time, by subtypes, hosts, and protein. IVDBVN also provides dead cases on total infected victims by time for countries.

With particular data to provinces, we could apply technology of Google map for displaying and following the distribution and spreading of influenza, so that managers and people could understand the distribution of influenza at provinces.

Moreover, the extraction of hidden information from sequences database will help researchers and managers get essential information, so that they could give out forecasts as well as preparation in respond with flu pandemic.

In addition, the particular information to provincial level of Vietnam could generate a friendly application with Vietnamese users, expressed through developing basement, interface and input data as well.

Developing information system for influenza to provinces of Asian countries. When making particular information to administration level under country, we could amend this product to natively using purpose.

Expanding more instruments, which permit users to carry out molecule biologic mathematic problems such as building phylogenetic tree, multiple alignments, finding BLAST.

ACKNOWLEDGMENT

The authors wish to acknowledge The Department of Science and Technology of Dong Nai Province, Vietnam that support the conditions to complete the works of own research group.

REFERENCES

[1] Dang Cao Cuong, Le Si Quang, Le Sy Vinh. *Influenza-specific amino acid substitution model*, the first international conference on knowledge DNA systems engineering, Hanoi, 2009.

[2] Bao Y., P. Bolotov, D. Dernovoy, et al,*The Influenza Virus Resource at the National Center for Biotechnology Information. J. Virol*. 2008 Jan; 82(2):596-601, 2008.

[3] Chang, S., Zhang, J., Liao, X., Zhu., et al.,*Influenza Virus Database (IVDB): an integrated information resource DNA analysis platform for influenza virus research. Nucleic Acids Res*, 35, D376-380, 2007

[4] Tien Dung Nguyen, The Vinh Nguyen, Dhanasekaran Vijaykrishna, et al.,*Multiple Sublineages of Influenza A Virus (H5N1), Vietnam, 2005-007*. Emerging Infectious Diseases 2008, Vol 14,632 – 636, 2008.

[5] Tran Van Lang, et al.,*Research for building informatics tools to process gene and protein's information*, Ministerial level thesis, Vietnam Academy of Sciences and Technology, 2004

[6] Tran Van Lang, *Using IT methods for solving molecular biology problems* (in Vietnam), Vietnam Education Publishing House, 2008, 230p.

[7] M.J. Zaki, C.J. Hsiao, *Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure*, IEEE Transactions on Knowledge and Data Engineering, 2005.

[8] Mohammed J. Zaki, Karam Gouda (2003), *Fast Vertical Mining Using Diffsets.*

[9] DNA Data Bank of Japan, http://www.ddbj.nig.ac.jp.

[10] Web API for Biology, http://xml.nig.ac.jp/wsdl/GetEntry.wsdl.