Published online 04 May 2015 *Nucleic Acids Research, 2015, Vol. 43, Web Server issue* ***W547–W551***
doi: 10.1093/nar/gkv417

MapMyFlu: visualizing spatio-temporal relationships between related influenza sequences

Nicholas Nolte[1,2], Nils Kurzawa[1,2], Roland Eils[1,2] and Carl Herrmann[1,2,*]

[1]Institute of Pharmacy and Molecular Biotechnology, and Bioquant Center, University of Heidelberg, Im Neuenheimer Feld 267, Heidelberg 69120, Germany and [2]Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, Heidelberg 69120, Germany

Received January 31, 2015; Revised April 10, 2015; Accepted April 18, 2015

ABSTRACT

Understanding the molecular dynamics of viral spreading is crucial for anticipating the epidemiological implications of disease outbreaks. In the case of influenza, reassortments or point mutations affect the adaption to new hosts or resistance to antiviral drugs and can determine whether a new strain will result in a pandemic infection or a less severe progression. To this end, tools integrating molecular information with epidemiological parameters are important to understand how molecular characteristics reflect in the infection dynamics. We present a new web tool, MapMyFlu, which allows to spatially and temporally display influenza viruses related to a query sequence on a Google Map based on BLAST results against the NCBI Influenza Database. Temporal and geographical trends appear clearly and may help in reconstructing the evolutionary history of a particular sequence. The tool is accessible through a web server, hence without the need for local installation. The website has an intuitive design and provides an easy-to-use service, and is available at http://mapmyflu.ipmb.uni-heidelberg.de

INTRODUCTION

With the increased mobility and globalization of the human population the threat of epidemic spreading is constantly raising, as illustrated dramatically by the recent Ebola outbreak. While not as lethal, influenza outbreaks still represent major health and economic threads. The avian flu outbreak in 2003 in southeast Asia resulted in economic losses of about 1 billion dollars, according to the Food and Agriculture Organization of the United Nations (FAO). To face these threats, better prediction and greater knowledge about influenza virus epidemiology are needed in order to preserve the worldwide health by improving prevention strategies. In particular, monitoring the dynamic of spreading and adaptation of the viral strain is important to anticipate possible evolutions, like resistance to anti-viral drugs or adaptation to new hosts (1). For example, very recently, several cases of human infections with avian influenza strains H7N9 have been reported in China (2,3). To tackle these questions, tools and databases are required to integrate various information sources, such as epidemiological characteristics and molecular signatures. Several influenza databases exist, such as the NCBI Influenza Virus Resource database (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html) (4), the Influenza Research Database (http://www.fludb.org/) (5) or the OpenFlu database (http://openflu.vital-it.ch) (6). The former two resources provide a number of analysis tools to perform multiple alignments or BLAST search, or identify relevant point mutations. On the other hand, other influenza related tools have been developed with a more epidemiological focus. Tools like Google Flu Trends (7), FluNearYou or FluTracking are already established and reveal flu occurrences using a crowd-sourcing approach, in the hope of anticipating evolution of influenza outbreaks. These tools estimate flu infections either by voluntary reportings or via internet queries related to influenza, with an accuracy that has been subject to debates (8,9). No genomic or proteomic information about the pathogen is included in these tools. An interesting tool combining numerous influenza data sources has recently been published (10). Here, sequence information for influenza strains obtained from the OpenFlu database is combined with disease information collected in the FAO EMPRES information system. A geographical visualization allows to relate disease characteristics to molecular events in the influenza strains. In this article, we present an online tool called MapMyFlu which allows the visualization of molecular relationships between influenza strains on a geographical map in a spatio-temporal fashion. The user can query a sequence of interest, for example from a newly sequenced strain, and obtain the list of related strains based on a BLAST query, geographically displayed according to the region where these strains were isolated and the time point of their collection. MapMyFlu performs a BLAST against a local instance of the NCBI influenza database and displays the most related hits using the Google Maps API. Temporal information about

*To whom correspondence should be addressed. Tel: +49 6221 423612; Fax: +49 6221 423612; Email: carl.herrmann@uni-heidelberg.de

the time point of virus isolation is given via an interactive histogram, which allows to display hits for a particular time period only. Different host species of the influenza virus obtained in the BLAST query are represented by different sets of icons, which can be selectively highlighted or hidden to focus on a particular type of carriers.

MATERIALS AND METHODS

User interface

The website created for MapMyFlu is kept simple with a focus on usability. The home page is composed of an input form for the query sequence together with some relevant parameters for the BLAST search. Given the high degree of similarity between influenza sequences, we use the percent similarity rather than the E-value as a criteria for selecting Blast hits. A minimal similarity can be specified, as well as maximum similarity, allowing to deal with cases in which a very large number of sequences are perfectly identical to the query sequence. The user can also restrict the search to a time period, specified in years. When submitting the form, a local BLAST request is performed against local instances of the NCBI influenza databases, using the BioPerl StandAloneBlast module. This allows an independent and more time-efficient use of MapMyFlu independent of the load of NCBI servers. For each hit, the accession number of the sequence is extracted and used to query a local SQLite database in order to retrieve relevant informations such as the name of the host species, the date of isolation of the sequence and the geographical location, which is then used to place the icons on the map. To represent the hit on the map, we have implemented four types of markers for host organisms, namely human, avian host, swine and others. The color of the marker (light to dark) represents the degree of similarity in the alignment (low to high). On the top of the page, a histogram displays the number of hits per year and per host species. Moving the mouse over the histogram displays marker from a specific year, giving an impression on the temporal evolution. Clicking on the legend buttons toggles markers associated with a particular host type. Each icon on the map can be clicked to display more detailed informations about the hit sequence in an information bubble. This bubble contains a link to the corresponding GenBank entry as well as the Influenza Research Database (IRD) entry, and details about the alignment (score, similarity, number of mutations, etc.). For a full analysis, the complete raw blast output is available as a text file by clicking on the appropriate button on the left of the result screen. Finally, an additional button on the output page gives access to a table summarizing all the hit sequences displayed along with the metadata associated with each of the sequences (host species, year of isolation, GPS coordinates, etc.). The output can be sorted according to different criteria by clicking on the column headers and sequences can be selected in the table using check boxes to download the corresponding FASTA sequences.

Technical implementation and maintenance

Besides the use of the BioPerl StandAlone Blast module, MapMyFlu is based on a SQLite database containing several tables: two tables (meta_aa and meta_na) that associate each protein/nucleotide Genbank accession number present in the local Blast database to metadata information, such as the year of isolation, the host category (human, avian, swine, others) and the country in which the sequence was isolated. These informations are directly obtained from a text file which is downloaded together with the sequence files from the NCBI Influenza ftp site during each update. In addition, these two tables contain a 'parsed_location' field which is obtained by parsing the fasta header of the sequences in the database. Often, but not always, the header contains a more precise geographical information than the country. For example, the sequence with accession number AFH00317 is associated with the country 'Japan', and its fasta header is A/Yamagata/56/1993(H3N2)), indicating the city of origin. Sometimes, the header only contains information about the country (Sequence AAA43373 : A/quail/Italy/1117/1965(H10N8)). In other cases, however, the fasta header is not formatted correctly and no useful geographical information can be extracted. For example, the sequence AAA43146 is associated with the country 'Japan', and its fasta header is A/duck/7/1982(H3)), which does not contain any information related to localization. In the latter two cases, only the country name is available. GPS coordinates are stored in a separate table of the database. This table associates a combination 'parsed_location,country' to GPS coordinates, to avoid ambiguities in the town name (for example we need to distinguish Hamburg, USA and Hamburg, Germany). The update of the Blast and SQLite databases is done automatically every three month. The update script adds information about new sequences to the meta_aa / meta_na tables, extracted from the information files downloaded along the sequence files. Additional information parsed from the fasta header (location, host, detailed date) are also added as additional fields to the tables.

The script then checks whether the combination 'parsed_location,country' is already registered in the table containing the GPS coordinates. Locations not yet represented are queried using the Google geocoding APIs, and GPS coordinates are then added to the table. All sequences present in the BLAST database can be associated with GPS coordinates, since as a last resort, the country name is used. In this case, the information bubble obtained by clicking the icons contains a warning to indicate that the placement of the icon is only based on country name. In general, the geocoding API reports coordinates of the center of the country when queried with only the country name. Errors can still occur, mainly when a header is incorrectly formatted and yields a wrong location when parsed. For example, in a previous version of our parser, the header A/duck/chicken/Nigeria/08RS848-20/2006(H5N1))
yielded besides 'Nigeria' the word 'chicken' as a geographical indication. However, the association 'Chicken,Nigeria' returns a valid GPS coordinate when submitted to the geocoding API. We have then modified the parser to take this into account and assign the location to 'Nigeria'. However, a perfect and automatic error correction is hard to implement. We therefore provide a button in the info bubble that can be clicked to report this kind of errors. An email is sent to the administrators containing the

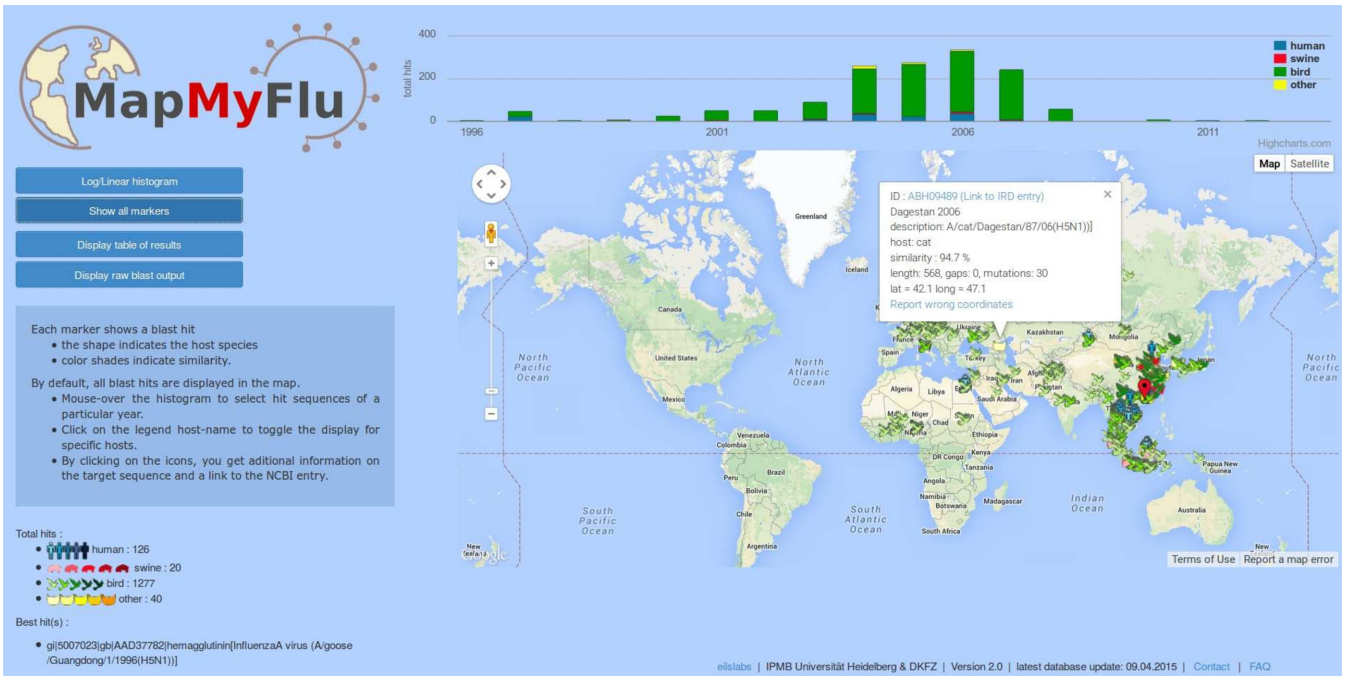

Figure 1. Result obtained by querying with the hemagglutinin sequence of a strain related to the 1996 H5N1 avian pandemic. The histogram displays the number of hits for each year, with a color indicating the host. The legend on the left gives the number of hits per host category. The shading of the icons indicate similarity, which in this case highlights that the most similar sequences are from southeast Asia. Several waves of infections are observable, and cases of human transmission in 1997 and after 2003 are also represented.

accession number of the faulty sequence. Older versions of the databases are saved in order to allow restoring previous queries on user demand. For an easy-to-use interface, the map is built using the Google Maps API and the histogram is built with Highcharts, a javascript library allowing to build interactive charts. Usage of Bootstrap and jQuery allow intuitive handling and instant reply of the website MapMyFlu.

RESULTS

First example: avian H5N1 influenza outbreak

To demonstrate the use of MapMyFlu, we first used a sequence from the 1996 H5N1 pandemia, which initiated in southeast China. The first reported and sequenced strain is a goose sequence (A/Goose/Guangdong/1/96(H5N1)). Inserting the hemagglutinin sequence of this strain into MapMyFlu yields a graphical output which recapitulates the major characteristics of the pandemia (11) (see Figure 1). First, a burst in 1996/1997 confined to the region where the first outbreak occurred in the Guandong province. In 1997, several human cases were reported in Hong Kong (12), which appear on the map. Second, a new and more pronounced outbreak starting in 2001, with several human cases again, in particular in Hong-Kong in 2003. Third, a transmission to different hosts starting with this second phase, in particular to swine, first in China, and from 2005 on to other areas in Asia like Indonesia. Taking the number of sequences present in the databases as a proxy of the number of infection cases, we see that the prevalence of H5N1 in swine is low, indicating a low swine-to-swine transmission rate (13). Given the role of pigs as virus reservoir intermediate between avian and human virus, this appearance of swine infections is an important characteristic of the disease progression. Lastly, the histogram clearly indicates a peak in the number of matching sequences around 2006/2007, corresponding to reports of a decline of infections after 2007. Hence, the MapMyFlu output recapitulates the temporal and geographical progression of the H5N1 influenza outbreak started in 1996.

Second example: H1N1 pandemia in 2009

As a second example the hemagglutinin sequence of an Influenza A virus (A/swine/Manitoba/SG1433/2009(H1N1)) is used to illustrate the outbreak of the influenza A H1N1, which started in 2009. By submitting this sequence to MapMyFlu with 1500 target sequences, the antigenic shift resulting in a transmission of the virus from swine to human in the year of the outbreak can be observed (Figure 2). Until the outbreak in 2009, the virus circulated mainly in swine hosts. After the host shift from swine to human in 2009, H1N1 infections in 214 countries from all over the world were reported and lead to 18,449 deaths (14). The first human cases were detected in North America and Mexico. MapMyFlu shows influenza sequences of pigs in the region around Canada in 2008 right before the pandemia started. These hits have only a minor number of mutations in comparison to the epidemic influenza in 2009, which confirms the initial epidemic outbreak in that region. The analysis of the neuraminidase sequence of the same virus shows that, except in few cases, the neuraminidase has remained confined to swine hosts.

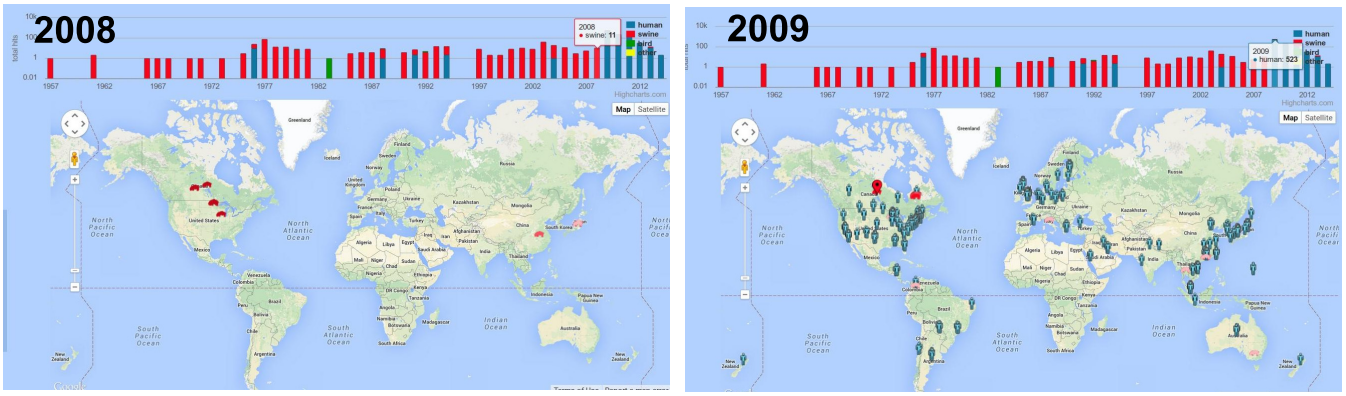

Figure 2. Evolution of the H1N1 2009 pandemic. The result of the MapMyFlu analysis on the hemagglutinin of a swine H1N1 strain shows how the transmission from swine to human occurred during 2009, leading to a burst of human infections around the globe. Note the logarithmic scaling of the histogram.

DISCUSSION

Proper visualization of complex and heterogeneous data sets such as molecular and epidemiological data is a key to the interpretation of the relationships between different components of a phenomenon. In the case of disease progression, how molecular changes in the sequence of the pathogen affect or are correlated to the speed of transmission, the virulence of the strain, or environmental parameters such as climatic conditions is still not completely understood. In this article, we have presented a tool, MapMyFlu, which attempts to represent in a very simple way molecular characteristics of influenza strains such as sequence similarities between strains as determined from sequence alignments, and geographical localization of these strains, together with informations about the host organism. To improve usability of the tool, we have used components such as the Google Maps API which are familiar to most users, and added interactive graphical charts to make the output the most interactive possible. MapMyFlu allows to simply track the history of a particular influenza sequence and immediately spot events such as rapid progression, progression to different geographical areas or transmission to new hosts. A limitation of the tool is that it is not yet coupled to epidemiological data sets or other disease databases nor to genetic information related to the virulence of the strain, based for example on specific sequence markers. Other webservers, databases or tools such as OpenFlu or EMPRES provide this type of data integration. We consider MapMyFlu as a complementary tool to these very comprehensive resources. However, it would be interesting to increase the complementarity between these tools; one possibility would be to overlay epidemiological informations related, for example, to disease outbreaks, onto the Google Maps representation of the Blast hits. So far, we simply provide links to IRD/fludb entries for the sequences identified in the Blast search. However, even in the absence of additional information, we have shown in the examples that the occurrences of sequences in the NCBI influenza database are a reasonable proxy for the importance of an influenza outbreak. Hence, we believe that using MapMyFlu will help in building further research hypothesis for researchers investigating molecular and epidemiological characteristics of influenza viruses.

ACKNOWLEDGEMENTS

We thank S. Marillet and L. Goetz for their pioneering work on this project. We also thank K.H. Groß for technical support in the implementation of the website. We thank the reviewers for their comments and suggestions to improve the tool.

FUNDING

Funding for open access charge: DKFZ internal funding.
Conflict of interest statement. None declared.

REFERENCES

1. Taubenberger,J.K. and Kash,J.C. (2010) Influenza virus evolution, host adaptation, and pandemic formation. *Cell Host Microbe*, **7**, 440–451.
2. Arunachalam,R. (2014) Adaptive evolution of a novel avian-origin influenza A/H7N9 virus. *Genomics*, **104**, 545–553.
3. Chen,F., Li,J., Sun,B., Zhang,H., Zhang,R., Yuan,J., Ou,X., Ye,W., Chen,J., Liu,Y. *et al.* (2011) Isolation and characteristic analysis of a novel strain H7N9 of avian influenza virus A from a patient with influenza-like symptoms in China. *Int. J. Infect. Dis.*, **33**, 130–131.
4. Bao,Y., Bolotov,P., Dernovoy,D., Kiryutin,B., Zaslavsky,L., Tatusova,T., Ostell,J. and Lipman,D. (2008) The influenza virus resource at the National Center for Biotechnology Information. *J. Virol.*, **82**, 596–601.
5. Squires,R.B., Noronha,J., Hunt,V., García-Sastre,A., Macken,C., Baumgarth,N., Suarez,D., Pickett,B.E., Zhang,Y., Larsen,C.N. *et al.* (2012) Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza Other Respir. Viruses*, **6**, 404–416.
6. Liechti,R., Gleizes,A., Kuznetsov,D., Bougueleret,L., Le Mercier,P., Bairoch,A. and Xenarios,I. (2010) OpenFluDB, a database for human and animal influenza virus. *Database (Oxford)*, **2010**, baq004.
7. Ginsberg,J., Mohebbi,M.H., Patel,R.S., Brammer,L., Smolinski,M.S. and Brilliant,L. (2009) Detecting influenza epidemics using search engine query data. *Nature*, **457**, 1012–1014.
8. Olson,D.R., Konty,K.J., Paladini,M., Viboud,C. and Simonsen,L. (2013) Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Comput. Biol.*, **9**, e1003256.
9. Valdivia,A., López-Alcalde,J., Vicente,M., Pichiule,M., Ruiz,M. and Ordobas,M. (2010) Monitoring influenza activity in Europe with

Google Flu Trends: comparison with the findings of sentinel physician networks - results for 2009-10. *Euro Surveill.*, **15**, PMID:20667303.
10. Claes,F., Kuznetsov,D., Liechti,R., VonDobschuetz,S., Truong,B.D., Gleizes,A., Conversa,D., Colonna,A., Demaio,E., Ramazzotto,S. *et al.* (2014) The EMPRES-i genetic module: a novel tool linking epidemiological outbreak information and genetic characteristics of influenza viruses. *Database (Oxford)*, **2014**, bau008.
11. Wan,X.F. (2012) Lessons from emergence of A/goose/Guangdong/1996-like H5N1 highly pathogenic avian influenza viruses and recent influenza surveillance efforts in southern China. *Zoonoses Public Health*, **59** (Suppl. 2), 32–42.
12. Chan,P.K.S. (2002) Outbreak of avian influenza A(H5N1) virus infection in Hong Kong in 1997. *Clin. Infect. Dis.*, **34**(Suppl. 2), S58–S64.
13. vanReeth,K. (2006) Avian influenza in swine: a threat for the human population? *Verh. K. Acad. Geneeskd. Belg.*, **68**, 81–101.
14. Cheng,V.C.C., To,K.K.W., Tse,H., Hung,I.F.N. and Yuen,K.Y. (2012) Two years after pandemic influenza A/2009/H1N1: what have we learned? *Clin. Microbiol. Rev.*, **25**, 223–263.