ISSN 1466-2523

Animal Health Research Reviews **11**(1); 73–79
doi:10.1017/S1466252310000083

Genome informatics of influenza A: from data sharing to shared analytical capabilities

Daniel A. Janies*, Igor O. Voronkin, Manirupa Das, Jori Hardman, Travis W. Treseder and Jonathon Studer

Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA

Received 4 March 2010; Accepted 26 March 2010

Abstract

Emerging infectious diseases are critical issues of public health and the economic and social stability of nations. As demonstrated by the international response to the severe acute respiratory syndrome (SARS) and influenza A, rapid genomic sequencing is a crucial tool to understand diseases that occur at the interface of human and animal populations. However, our ability to make sense of sequence data lags behind our ability to acquire the data. The potential of sequence data on pathogens is not fully realized until raw data are translated into public health intelligence. Sequencing technologies have become highly mechanized. If the political will for data sharing remains strong, the frontier for progress in emerging infectious diseases will be in analysis of sequence data and translation of results into better public health science and policy. For example, applying analytical tools such as Supramap (http://supramap.osu.edu) to genomic data for pathogens, public health scientists can track specific mutations in pathogens that confer the ability to infect humans or resist drugs. The results produced by the Supramap application are compelling visualizations of pathogen lineages and features mapped into geographic information systems that can be used to test hypotheses and to follow the spread of diseases across geography and hosts and communicate the results to a wide audience.

Keywords: genomics, informatics, pathogens, phylogenetics, data sharing, GISAID, GenBank, Supramap, influenza A, epidemiology, evolution, public health

Non-sharing of data: 1997–2005

Avian influenza (H5N1) has been a concern since 1997, when it began to infect humans in Hong Kong. H5N1 has spread from its origins in China to other regions in Asia, Russia, India, Pakistan, the Middle East, North and West Africa, and Europe (Chen *et al*., 2006; Janies *et al*., 2007; Hovmöller *et al*., 2010). Since 2003 there have been 489 human cases and 289 deaths attributed to H5N1 (WHO, 2009a).

The research environment has changed dramatically since 1997, when an outbreak of avian influenza (H5N1) among poultry and humans in Hong Kong alerted health officials to its dangers. In the early 2000s, genomic sequencing became highly mechanized and computational power began to improve dramatically via Beowulf clustering (Gee, 2000). However, in the same period, genetic sequence data on avian influenza was not widely shared. Some data could be found in GenBank (http://ncbi.nlm.nih.gov), which is public domain. In parallel, a small group of influenza researchers (Nature, 2006) had access to a private database maintained on behalf of the World Health Organization (WHO) at Los Alamos National Laboratory (LANL) (http://flu.lanl.gov). The LANL database included data from GenBank, as well as private data submitted by WHO directly to LANL (Macken *et al*., 2001). Furthermore, common practice by most biomedical researchers, including the influenza community, was to share data only after publication. In contrast, at the turn of the millennium, large publicly funded projects such as the Human Genome Project were

*Corresponding author. E-mail: Daniel.Janies@osumc.edu

charged with posting data rapidly after collection (Nature, 2006).

In 2007, the International Health Regulations were revised, thus requiring nations to assess and report disease outbreaks to the WHO within 48 h (Wilson *et al.*, 2008). Previously, governments were not willing to share information about diseases within their borders, as demonstrated in the severe acute respiratory syndrome (SARS) epidemic (FlorCruz, 2003). Each person and family affected by SARS represents a personal tragedy. At the level of international trade, perceptions of public health in a country can lead to economic tragedy. Estimates for the cost of SARS, in terms of lost productivity, are US $18 billion for East and Southeast Asian economies (ADB, 2003). If there had been ongoing public concern, the cost of SARS in Asia could have been US $60 billion or higher (ADB, 2003). SARS cost the Canadian economy an estimated CDN $519 million in lost revenues in 2003 (CBC, 2003). These losses are very high compared to the actual scale of the epidemic, representing over US $2 million per person infected in 2002–2003.

Along with influenza, the SARS epidemic represented an important event requiring changes to public health policy. SARS made it clear that novel infectious diseases can start in one country and quickly spread throughout the whole world. In light of SARS and influenza, many in public health informatics began to realize that data sharing and cooperation across the globe are vitally important to prepare for and respond to emergent pathogens (Nature, 2006). The CDC predicted in the United States alone that an H5N1 pandemic among humans would affect 15–35% of the population, resulting in $70–167 billion in lost productivity (Gerberding, 2005).

The severity of the H1N1 pandemic that began in early 2009 appears to be much lower than projected (Presanis *et al.*, 2009). As of 19 March 2010, pandemic H1N1 has caused at least 16,813 deaths worldwide (WHO, 2010). These deaths are heartbreaking for the families involved. Furthermore, H1N1 caused fear and absenteeism in some regions of the world such as Mexico (Stevenson, 2009), Ukraine (Pan, 2009), and Argentina (Valente, 2009).

The economic aspects of the current H1N1 pandemic are still being determined. In the United States, any economic impacts of H1N1 will have been overshadowed by the trillions of dollars of costs due to the banking crisis and recession that began in 2007 and has continued into 2010 (Barr, 2009).

There are arguments, at both the international and individual levels, that data on infectious diseases should not be shared as some might profit from the data collected by others. For example, leaders in Indonesia have argued that isolates of avian influenza (H5N1) isolated in their country represent biological patrimony (Garrett and Fidler, 2007). As such, Indonesia asked for a guarantee that any vaccine developed based on influenza isolates from their country should be in turn shared with their citizens (Wilson *et al.*, 2008). In a separate example, China did not share data because Western researchers did not give credit to Chinese researchers who collected the primary data (Zamiska, 2006).

Data sharing: 2006–2009

The tradition in research has been that data are private until published (Brown, 2006). However, in light of the public health concerns many have realized that this practice inhibits innovation (Bogner *et al.*, 2006). Data sharing is imperative for progress. When data are shared, researchers across many different disciplines can attack a problem. If data access is restricted to individuals well versed in established technologies and schools of thought, then the next generation of thinkers is limited in its ability to bring new technologies and approaches to the problem. Pandemic influenza presents challenges not easily met by molecular or cellular biology or epidemiology alone. Molecular and cellular biologists are concerned with the interactions of the host's and pathogens' genome(s). Epidemiologists are concerned with the occurrence of diseases and symptoms. Epidemiologists rarely consult genomic data, but rather count classes of symptoms such as 'influenza-like illnesses'. Neither discipline has been able to develop an understanding of an outbreak of disease from both genomic and epidemiological points of view. Collaboration is necessary for a bridge between molecular biology and epidemiology to be built. For example, virologists, molecular biologists, evolutionary biologists, geographers, and computer scientists have collaborated to combine genomic sequences with geographic and temporal information. These collaborations have provided maps of the movement of pathogens and key mutations among pathogens, such as those that confer resistance to drugs or the ability to jump from animals to humans (Janies *et al.*, 2007, 2008, 2010a, b; Hill *et al.*, 2009; Hovmöller *et al.*, 2010). The next step is to turn genomic data and maps into forecasts of the spread of specific pathogens or mutations. We have taken a first step in this direction by using maps, phylogenetics, and studies of selective pressure for resistance to oseltamivir (Tamiflu®) on seasonal H1N1 lineages to develop early and spatially specific warning of the rise of resistance to oseltamivir in pandemic H1N1 lineages (Janies *et al.*, 2010b). Further incorporation of genomic and geographic data into pathogen surveillance and public health intelligence will require data sharing as well as interaction among molecular and evolutionary biologists, mathematicians, statisticians, geographers, and public health scientists.

One bright spot in the area of data sharing has been the Influenza Genome Sequencing Project (IGSP) (http://www3.niaid.nih.gov/LabsAndResources/resources/gsc/Influenza). Since its origins in 2005, the IGSP has sequenced and released over 4650 influenza genomes

https://doi.org/10.1017/S1466252310000083 Published online by Cambridge University Press

into the public domain, pre-publication. The IGSP is under contract from the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH) of the United States to sequence genomes of viral isolates, collected by over 40 contributing laboratories around the world, and to rapidly put them in the public domain, via a special section of GenBank called the Influenza Virus Resource (Bao *et al.*, 2008). This practice is good for the research community at large but has hindered the IGSP from collaborating with those who prefer to maintain right to publish before data are released. In contrast, other groups have been successful in this model of data release and rapid publication. For example, Salzberg *et al.* (2007) released data first and then published as an international team. Authorship credit was shared widely among the team who played various roles such as: providing the isolates, sequencing the isolates, analyzing the data, and writing the paper.

The H1N1 pandemic that began in 2009 has brought several challenges and has evoked renewed responses by the research community for data sharing. Foremost among the responses is the Global Initiative for Sharing of All Influenza Data (GISAID). Due to the efforts of GISAID, the Centers for Disease Control (CDC) in the United States, and the WHO, sequence data for the 2009 pandemic strain of H1N1 were released swiftly into a publicly accessible database on 25 April 2009 (the EpiFluDB^{TM} database section of gisaid.org), pre-publication (WHO, 2009b). Large amounts of data also flowed rapidly into GenBank as of 27 April 2009 (e.g. http://www.ncbi.nlm.nih.gov/genomes/FLU/SwineFlu2009.html).

Studies on these early sequences were rapidly produced to examine the host and origins of pandemic H1N1 (Garten *et al.*, 2009; Novel Swine-Origin Influenza A H1N1 Virus Investigation Team, 2009). The actions of GISAID/CDC/WHO and the IGSP are watershed events for infectious disease genomics and public health that hopefully have a lasting impact on attitudes for data sharing.

As of late April 2009, no longer did the world have to wait for any particular scientist or organization to decide that their H1N1 data were ready to be shared. Public health organizations around the world followed the lead of GISAID/CDC/WHO and IGSP and quickly began to submit pre-publication sequence data to GISAID and GenBank. Currently, GISAID has the largest collection of sequence data for the H1N1 pandemic lineage, and there is little to no unique data in Genbank. However, at the same time, sequence data from other strains and the IGSP have continued to be submitted to NIH's GenBank. These multiple data sources can create a challenge for the researcher who is concerned with comprehensive sampling of influenza data. For example, in a phylogenetic analysis, the sampling of the sequence diversity is very important to increase the accuracy of the study (Zwickl and Hillis, 2002). In our laboratory, we found it useful to check various databases and create a set of non-overlapping sequences. Our solution is presented in the Methods section below. Recently, GISAID started to mirror sequence data from GenBank and other repositories. This mirroring has abated our concerns of not having a complete dataset. However, the speed of mirroring is very important. For those public health scientists charged with disease surveillance, having timely access to sequence data is critical. Despite all the progress in data sharing, GISAID's independence was temporarily in peril in the summer of 2009, requiring the Max-Planck-Institute for Informatics (MPI) in Saarbrücken, Germany, to step in as the new host for the EpiFlu^{TM} database, following a challenge over its control by the Swiss Institute of Bioinformatics (SIB) (Butler, 2009a, b). In collaboration with GISAID's user community, WHO Collaborating Centers for Influenza, and veterinary reference laboratories of the Food and Agriculture Organization (FAO) of the United Nations and the World Organization for Animal Health (OIE), MPI promptly developed and offered users a replacement to SIB software. Subsequently, in October 2009, the German Government announced it had entered into an agreement with GISAID to become the long-term host of its EpiFlu^{TM} database in Bonn, from 2011 onwards (Federal Ministry of Food, Agriculture and Consumer Protection, Germany, 2009). The agreement also calls for the curation of GISAID data by the Friedrich-Löffler-Institute, the national research center for animal health of Germany, to improve the overall quality of data offered to GISAID users.

Other issues adding complexity for researchers are the various data access agreements of GISAID and GenBank. GenBank, the DNA Data Bank of Japan (DDBJ), and the European Molecular Biology Laboratory (EMBL) form the International Nucleotide Sequence Database Collaboration (INSD) (http://www.insdc.org). Highlights of the International Nucleotide Sequence Database Collaboration Policy (Brunak *et al.*, 2002) include: 'The INSD has a uniform policy of free and unrestricted access to all of the data records their databases contain' and 'The INSD will not attach statements to records that restrict access to the data, limit the use of the information in these records, or prohibit certain types of publications based on these records. Specifically, no use restrictions or licensing requirements will be included in any sequence data records, and no restrictions or licensing fees will be placed on the redistribution or use of the database by any party.'

In contrast, a GISAID user must sign an agreement to become an authorized user (GISAID, 2008). A passage of this agreement states: 'Distribution of Data. You agree that Data may be provided to all GISAID Data users and providers that have agreed to be bound by this Agreement (or an Agreement with substantially similar terms) and continue to abide by its terms (collectively "Authorized Users"). However, subject to applicable law, You agree not to distribute Data to any third party other than Authorized Users, except for Data that has been expressly and lawfully placed in the public domain by GISAID in

https://doi.org/10.1017/S1466252310000083 Published online by Cambridge University Press

accordance with subparagraph (a), by You, or by any other party having the right to do so.'

The key distinction is that GenBank data can be repurposed and distributed by anyone at will. GISAID data can only be repurposed and distributed to GISAID users who sign the agreement. This distinction on how data can be repurposed has important motivations and implications. GISAID intends the user agreement to be a vehicle to enable data sharing by allowing access to those users who agree to collaborate with and acknowledge data providers. Acknowledgement of data providers is something that is not often done in the context of use of data from GenBank, other than citation of the accession numbers for the sequences used. Irrespective of the database from which a sequence was drawn, the basic scholarship and public policy (treated below) that calls for citation of papers can easily be extended to the acknowledgment of the originating laboratories and scientists that provided the isolates, sequence data, and metadata. This acknowledgement can be accomplished via a table of supplemental data (e.g. http://www.ij-healthgeographics.com/qc/content/9/1/13/suppl/S9) or co-authorship (e.g. Salzberg *et al.*, 2007). Common practice among scientists who publish results based on sequence data, such as posting a multiple sequence alignment containing GISAID data, can be done in the 'wiki' section, which is in the password protected section of GISAID.org, without violating the agreement. In contrast, an alignment of GenBank data can be posted openly (e.g. http://supramap.osu.edu/sm/supramap/publications). The practice of making raw data and research results that contain raw data public after publication is important as it allows any scientist to repeat the experiments. For NIH-funded investigators in the United States, the practices of timely sharing of data and acknowledgement of sources are codified (NIH, 2003). Furthermore, timely public dissemination (within a year) of papers based on NIH-funded research has recently been made a law in the United States (NIH, 2008). Similar policies have been enacted or considered around the world (Alliance for Taxpayer Access, 2006).

All of the analytical applications on which we have published in the peer-reviewed literature (Janies *et al.*, 2010a, http://supramap.osu.edu; Hovmöller *et al.*, 2010, http://routemap.osu.edu; and Janies *et al.*, 2010b, http://pointmap.osu.edu) are available freely over the web. These applications do not provide data. They are analytical and visualization tools to which users must supply datasets. However, to some users, data security is very important and a web-based application is unsuitable. In these cases, users can compile and run a binary of POY (Varón *et al.*, 2009) enabled for Supramap analysis on computers in a local secure environment without the use of a remote cluster and without transmission of their data outside of their organization. We include instructions for the stand-alone application here: http://supramap.osu.edu/sm/supramap/tutorials#section2.

Another area for improvement in influenza informatics is sharing of metadata such as host, place of isolation, date of isolation, and clinical information (Janies *et al.*, 2007; Butler, 2008). Metadata standards have improved in both GISAID and GenBank. Binomial names are now being used for many host species (e.g. *Anas crecca*, *Anas platyrhynchos*, *Anas boschas*, *Anas clypeata*, and *Anas acuta*). Significant improvements in GISAID include important fields for metadata of animal hosts 'domestic (or wild) status', 'health status', 'vaccination status', and 'specimen source (type of tissue or swab)'. Nevertheless, there is still much room for improvement in the data that providers submit. For example, in a recent study, metadata associated with NA sequences for H1N1 data contained many common names for host species (e.g. 'duck', 'goose', 'swan', and 'chicken') as well as some unhelpful names for hosts (e.g. 'other avian' and 'other mammal').

There are also other databases that specialize in phenotype prediction and vaccine design. The Influenza Research Database (fludb.org) includes drug resistance and virulence genotypes, information on epitope variation important for vaccine design, clinical metadata, and surveillance information. GISAID provides discrete metadata required by the collaborating centers of the WHO that are part of the vaccine strain selection process. This has permitted the raw data of candidate strains to be available to GISAID users, but their candidacy for a vaccine is unknown, except to the collaborating centers.

Beyond data sharing: 2010 to the future

Large-scale sequencing and rapid global sharing of influenza genomes have been the breakthrough events of the past decade. However, raw sequence data cannot, on their own, provide the information needed by public health officials. Public health scientists must integrate knowledge of the genomes of pathogens with host biology as well as societal and environmental factors to understand the etiology epidemics and to anticipate their trajectories. One of the most effective ways to integrate diverse information is to put the data together in a geographic information system (GIS) (Fig. 1). A GIS provides an interdisciplinary framework for hypothesis generation and testing and can be used to communicate results. To these ends, we have developed several mapping applications to integrate genomic data with geographic and phenotypic information in interactive visual environments that allow public health scientists and policy makers to determine: (i) the number and type of distinct pathogen strains that circulate in a region (Janies *et al.*, 2007); (ii) from which geographic sources and hosts, the pathogens originate (Fig. 1) (Janies *et al.*, 2007, 2008; Hovmöller *et al.*, 2010; http://routemap.osu.edu); and (iii) whether these strains are becoming resistant to drugs (Hill *et al.*, 2009; Janies *et al.*, 2010b; http://pointmap.osu.edu) or adapting from avian to mammalian hosts

https://doi.org/10.1017/S1466252310000083 Published online by Cambridge University Press

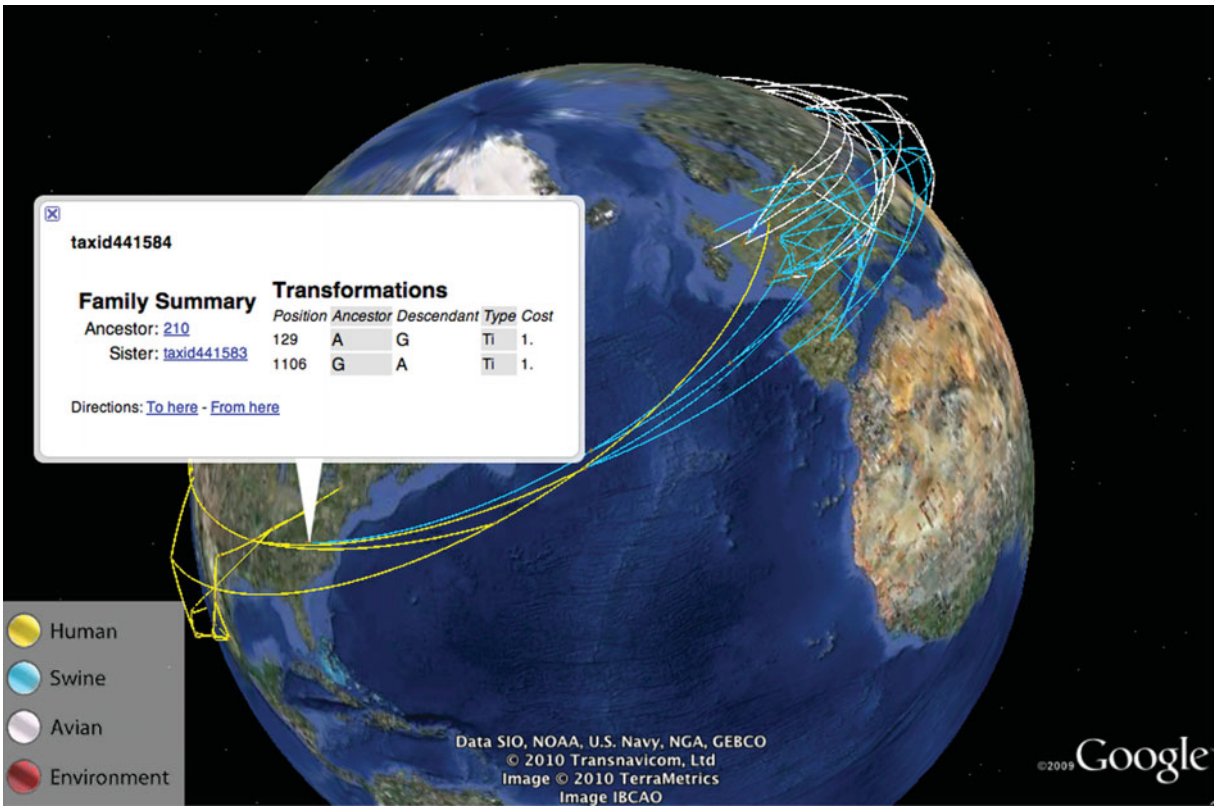

Fig. 1. Screen capture of an interactive phylogenetic map of the emergence in early 2009 of the pandemic lineage of H1N1 influenza. The tree is based on nucleotide sequences for the neuraminidase segment. The colors of branches of the tree indicate the sample (animal host or environment) from which the virus was isolated. Mutations at each node of the tree can be viewed in pop-up windows. This phylogenetic map was created with Supramap software (Janies *et al.,* 2010a) and is available for download as a keyhole markup file (KML) compatible with Google Earth™ at http://supramap.osu.edu/sm/supramap/publications.

(Janies *et al.*, 2007, 2010a; http://supramap.osu.edu). The major advantage of genetic and phenotypic maps over syndromic approaches (Brownstein *et al.*, 2008) is that genetic and phenotypic maps allow public health officials to make decisions based on the biology of specific pathogens and hosts, rather than the occurrence of symptoms, which can arise from unrelated diseases.

Conclusions

Sequencing and rapid data sharing seem to be on a strong footing in influenza. The next step is to share analytical and visualization tools to foster interactions among many disciplines. Our aims include reaching out to diverse groups of students and researchers, public health scientists, and policy makers. The best way to ensure that these outreach efforts are successful is to make the tools as user-friendly and accessible to the raw data as possible.

Methods

Data sharing has led to a proliferation of data sources and large amounts of sequence data and metadata of various levels of quality for influenza. In order to carry out phylogenetic and biogeographic analyses in our laboratory, we find it necessary to federate data sources and verify the quality of sequences and metadata. We built a custom database application populated with a non-overlapping set of sequence data from GenBank and GISAID and further annotated the data with geographical information such as latitude and longitude.

In terms of quality controls, we removed sequences that were difficult to align due to their short length or mutations that broke the reading frame of the alignment. In the case of H1N1 data, we also verified annotation regarding whether a sequence is in H1N1 pandemic or seasonal lineages via alignment and phylogenetic analysis.

We take the following steps to integrate sequence and metadata from GenBank and GISAID. We download sequence and metadata data via standard graphical user interfaces provided by GenBank. (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html or http://ww.ncbi.nlm.nih.gov/sites/batchentrez) and GISAID (EpiFlu™ database).

For each isolate identifier in sequence data from GISAID (EPI_ISOLATE_ID), we cross-referenced the

https://doi.org/10.1017/S1466252310000083 Published online by Cambridge University Press

accession numbers and taxids from GenBank. Many of the headers in the GISAID FASTA sequence file had a GenBank accession associated with them. We parsed the GISAID sequence file headers to obtain the GenBank accession, ran the accessions through GenBank's batch entrez tool, found taxids, and updated the data with the taxid values. We then searched GenBank for corresponding datasets. We used a Ruby script to parse the file and obtain metadata (taxid, accession, location, host, date, and strain name).

Once a CSV file was created with all available metadata, we used the UNIX commands (grep -v) and the taxid field of the GISAID data to filter out sequences that were duplicated between GISAID and GenBank. When duplicates were found we kept the GISAID record.

Other efforts allow for output of latitude and longitude information for isolates (see MacDonald *et al.*, 2009 for GenBank data). We have added latitude and longitude data for GenBank and GISAID data. For georeferencing, we developed a Ruby (http://www.ruby-lang.org) script that uses the location data and the metadata as input. We were able to put most location data in a hierarchical format (e.g. Columbus, Ohio, USA). The hierarchical format helps us to disambiguate locations with the same name (e.g. Paris, Texas versus Paris, France). Our Ruby script parses the hierarchical location data and queries the Geonames database (http://ipinfodb.com/ip_database.php) for latitude and longitude. We then checked the latitude and longitude data and added it to the metadata for each isolate.

Our database application is built using the Ruby on Rails (http://rubyonrails.org) web framework and the MySQL (http://www.mysql.com) database package. Our database application uses Rails Object-relational Mapping (ORM) to allow the user interface to leverage the database model. ORM also allows one to build and run queries to any field within the database, making it easy to customize and adapt the user interface based on changing requirements and other data federation problems. We provide a means for users to manage their own projects. Within a project, a user can create, run, and save various queries to the database. The user can access these queries later to re-run and or modify. A single query made to the database may consist of parameters such as the location (based on continent, country, or region), strain name, host, specific or all genomic segment(s), H1N1 lineage information, and a range of dates of collection. The system then provides three types of result: (i) nucleotides in FASTA format; the labels of the FASTA file are GISAID or GenBank accession numbers if it is a sequence unique to GenBank; (ii) a file of geographic and temporal references (decimal latitude, longitude, and date of isolation for each accession) in comma-separated values (CSV) format; and (iii) a metadata file that contains other information regarding the sequences such as gene segment, host, strain name, and location in tabular format. The nucleotide (FASTA) and geographic and temporal data (CSV) files are compatible with other systems for analysis such as Supramap (Janies *et al.*, 2010a; http://supramap.osu.edu).

Please contact us if you would like to know more about the custom database and other applications we discuss in this review. The infrastructure can be used to federate primary data with public data for ongoing projects within natural science and medical research.

Acknowledgments

We acknowledge the Department of Biomedical Informatics of the Ohio State University for providing space, funding, and administrative support. We thank the Medical Center Information Services team of OSU and the Ohio Supercomputer Center for hosting computing clusters used in this study. We acknowledge that this article is based upon work supported by, or supported in part by, the US Army Research Laboratory and Office under grant numbers W911NF-05-1-0271 and HR-0011-09-2-0009.

References

Asian Development Bank (ADB) (2003). http://www.adb.org/Documents/Books/ADO/2003/update/sars.pdf

Alliance for Taxpayer Access (2006). Worldwide momentum for public access to publicly funded research. Available online at http://www.taxpayeraccess.org/issues/access/access_resources/worldwide-momentum-for-public-access-to-publ.shtml

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. *Journal of Virology* **82**: 596–601.

Barr C (2009). Bank bailout could cost $4 trillion. Available online at http://money.cnn.com/2009/01/27/news/bigger.bailout.fortune/

Bogner P, Capua I, Lipman J, Cox NJ, Lipmans DJ et al. (2006). A global initiative on sharing avian flu data. *Nature* **442**: 981.

Brown D (2006). Bird flu fears ignite debate on scientists' sharing of data. Available online at http://www.washingtonpost.com/wp-dyn/content/article/2006/05/24/AR2006052402293.html

Brownstein J, Freifeld C, Reis B and Mandl K (2008). Surveillance sans frontières: internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Medicine* **5**: e151. doi:10.1371/journal.pmed.0050151.

Brunak S, Danchin A, Hattori M, Nakamura H, Shinozaki K, Matise T and Preuss D (2002). Nucleotide sequence database policies. *Science* **298**: 1333.

Butler D (2008). Politically correct names given to flu viruses. Available online at www.nature.com/uidfinder/10.1038/452923a

Butler D (2009a). Flu database rocked by legal row. *Nature* **460**: 786–787.

Butler D (2009b). Flu database row escalates. Available online at http://blogs.nature.com/news/thegreatbeyond/2009/09/flu_database_row_escalates.html

Canadian Broadcasting Corporation (CRC). (2003). The Economic Impact of SARS. Available online at www.cbc.ca/news/background/sars/economicimpact.htm

Chen H, Smith G, Li K, Wang J and Fan X (2006). Establishment of multiple sublineages of H5N1 influenza virus in Asia: implications for pandemic control. *Proceedings of the National Academy of Sciences, USA* **103**: 2845–2850.

Federal Ministry of Food, Agriculture and Consumer Protection, Germany (2009). http://www.bmelv.de/cln_102/SharedDocs/Pressemitteilungen/2009/249-Ai-Influenza-Datenbank%20Bonn.html

FlorCruz J (2003). China censors CNN SARS report. Available online at h ttp://www.cnn.com/2003/WORLD/asiapcf/east/05/14/sars.censor/

Garrett L and Fidler D (2007). Sharing H5N1 viruses to stop a global influenza pandemic. *Public Library of Science, Medicine* **4**: e330. doi:10.1371/journal.pmed.0040330.

Garten R, Davis T, Russell C, Shu B, Lindstrom S, Balish A, Sessions WM, Xu X, Skepner E, Deyde V, Okomo-Adhiambo M, Gubareva L, Barnes J, Smith CB, Emery SL, Hillman MJ, Rivailler P, Smagala J, de Graaf M, Burke DF, Fouchier RAM, Pappas C, Alpuche-Aranda CM, López-Gatell H, Olivera H, López I, Myers CA, Faix D, Blair PJ, Yu C, Keene KM, Dotson Jr PD, Boxrud D, Sambol AR, Abid SH, St. GeorgeK, Bannerman T, Moore AL, Stringer DJ, Blevins P, Demmler-Harrison GJ, Ginsberg M, Kriner P, Waterman S, Smole S, Guevara HF, Belongia EA, Clark PA, Beatrice ST, Donis R, Katz J, Finelli L, Bridges CB, Shaw M, Jernigan DB, Uyeki DM, Smith DJ, Klimov AI and Cox NJ (2009). Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science* **325**: 197–201.

Gee H (2000). Homegrown computer roots out phylogenetic networks. *Nature*. http://www.nature.com/nature/journal/v404/n6775/pdf/404214b0.pdf

Gerberding J (2005). Pandemic planning and preparedness. Hearing of the108th United States Congress Committee on Energy and Commerce.

GISAID (2008). GISAID EpiFlu Database Access Agreement. http://tinyurl.com/gisaid

Hill A, Guralnick R, Wilson M, Habib F and Janies D (2009). Evolution of drug resistance in multiple distinct lineages of H5N1 avian influenza. *Infection, Genetics, and Evolution* **9**: 169–178.

Hovmöller R, Alexandrov B, Hardman J and Janies D (2010). Tracking the geographic spread of avian influenza (H5N1) with multiple phylogenetic trees. *Cladistics* **26**: 1–13.

Janies D, Hill A, Guralnick R, Habib F, Waltari E and Wheeler WC (2007). Genomic analysis and geographic visualization of the spread of avian influenza (H5N1). *Systematic Biology* **56**: 321–329.

Janies D, Habib F, Alexandrov B, Hill A, Pol D (2008). Evolution of genomes, host shifts, and geographic spread of SARS-CoV and related coronaviruses. *Cladistics* **24**: 111–130.

Janies D, Treseder T, Alexandrov B, Habib F, Chen J, Ferreira R, Çatalyürek U, Varón A and Wheeler WC (2010a). The Supramap project: linking pathogen genomes with geography to fight emergent infectious diseases. *Cladistics*. In press.

Janies D, Voronkin I, Studer J, Hardman J, Alexandrov B, Treseder T and Valson C (2010b). Selection for resistance to oseltamivir in seasonal and pandemic H1N1 influenza and widespread co-circulation of the lineages. *International Journal of Health Geographics* **9**: 13.

MacDonald N, Parks D, Beiko R (2009). SeqMonitor: influenza analysis pipeline and visualization. *PLoS Current Influenza* 2009 September 22: RRN1040.

Macken C, Lu H, Goodman J and Boykin L (2001). The value of a database in surveillance and vaccine selection. In: Osterhaus ADME, Cox N and Hampson AW (eds) *Options for the Control of Influenza IV*, Vol. **1219**. Amsterdam: Elsevier Science, pp. 103–106.

Nature (2006). Dreams of flu data. *Nature* **440**: 255–256.

NIH (2003). Data Sharing Policy and Implementation Guidance. Available online at http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

NIH (2008). National Institutes of Health Public Access. Available online at http://publicaccess.nih.gov/index.htm

Novel Swine-Origin Influenza A H1N1 Virus Investigation Team (2009). Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *New England Journal of Medicine* **360**: 2605–2615.

Pan P (2009). In Ukraine, H1N1 pandemic sets off panic and politicking. Available online at http://www.washingtonpost.com/wp-dyn/content/article/2009/11/20/AR2009112004023_pf.html

Presanis A, De Angelis D, The New York City Swine Flu Investigation Team, Hagy A, Reed C, Riley S, Cooper B, Finelli L, Biedrzycki P and Lipsitch M (2009). The severity of pandemic H1N1 Influenza in the United States from April to July 2009: A Bayesian Analysis. *PLoS Medicine* **6**: e1000207. doi:10.1371/journal.pmed.1000207

Salzberg S, Kingsford C, Cattoli G, Spiro D, Janies D, Aly M, Brown I, Couacy-Hymann E, De Mia G, Dung D, Guercio A, Joannis T, Ali A, Osmani A, Padalino I, Saad M, Savić V, Sengamalay N, Yingst S, Zaborsky J, Zorman-Rojs O, Ghedin E, Capua I (2007). Genome analysis linking recent European and African influenza (H5N1) viruses. *Emerging Infectious Diseases* **13**: 5.

Stevenson M (2009). Mexico swine flu deaths spur global epidemic fears. Available online at http://www.guardian.co.uk/world/feedarticle/8473318

Valente M (2009). ARGENTINA: Experts Put H1N1 Flu Outbreak in Perspective. Available online at http://www.ipsnews.net/news.asp?idnews=47388

Varón A, Vinh L, Wheeler WC (2009). POY version 4: phylogenetic analysis using dynamic homologies. *Cladistics* **26**: 72–85.

WHO (2009a). Cumulative number of confirmed human cases of avian influenza A (H5N1) reported to WHO. Available online at http://www.who.int/csr/disease/avian_influenza/country/cases_table_2010_03_16/en/index.html

WHO (2009b). Viral gene sequences to assist update diagnostics for swine influenza A (H1N1). 25 April 2009. Available online at http://www.who.int/csr/disease/swineflu/Gene_sequences_20090425.pdf

WHO (2010). Pandemic (H1N1) 2009 – update 92. Available online at http://www.who.int/csr/don/2010_03_19/en/index.html

Wilson K, Tigerstrom B, and McDougall C (2008). Protecting global health security through the international health regulations: requirements and challenges. *Canadian Medical Association Journal* **179**: 44–48.

Zamiska N (2006). How academic flap hurt world effort on Chinese bird flu. *The Wall Street Journal*, 24 February 2006: A1.

Zwickl D and Hillis D (2002). Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology* **51**: 588–598.