Knowledge-Driven Methods for Geographic Information Extraction in the Biomedical Domain

by

Tasnia Tahsin

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2019 by the
Graduate Supervisory Committee:

Graciela Gonzalez, Co-Chair
Matthew Scotch, Co-Chair
George Runger

ARIZONA STATE UNIVERSITY

December 2019

ABSTRACT

Accounting for over a third of all emerging and re-emerging infections, viruses represent a major public health threat, which researchers and epidemiologists across the world have been attempting to contain for decades. Recently, genomics-based surveillance of viruses through methods such as virus phylogeography has grown into a popular tool for infectious disease monitoring. When conducting such surveillance studies, researchers need to manually retrieve geographic metadata denoting the location of infected host (LOIH) of viruses from public sequence databases such as GenBank and any publication related to their study. The large volume of semi-structured and unstructured information that must be reviewed for this task, along with the ambiguity of geographic locations, make it especially challenging. Prior work has demonstrated that the majority of GenBank records lack sufficient geographic granularity concerning the LOIH of viruses. As a result, reviewing full-text publications is often necessary for conducting in-depth analysis of virus migration, which can be a very time-consuming process. Moreover, integrating geographic metadata pertaining to the LOIH of viruses from different sources, including different fields in GenBank records as well as full-text publications, and normalizing the integrated metadata to unique identifiers for subsequent analysis, are also challenging tasks, often requiring expert domain knowledge. Therefore, automated information extraction (IE) methods could help significantly accelerate this process, positively impacting public health research. However, very few research studies have attempted the use of IE methods in this domain.

This work explores the use of novel knowledge-driven geographic IE heuristics for extracting, integrating, and normalizing the LOIH of viruses based on information available in GenBank and related publications; when evaluated on manually annotated test sets, the methods were found to have a high accuracy and shown to be adequate for addressing this challenging problem. It also presents GeoBoost, a pioneering software system for georeferencing GenBank records, as well as a large-scale database containing over two million virus GenBank records georeferenced using the algorithms introduced here. The methods, database and software developed here could help support diverse public health domains focusing on sequence-informed virus surveillance, thereby enhancing existing platforms for controlling and containing disease outbreaks.

DEDICATION

To my mother, whose life-long dedication to family and education
painstakingly paved out my path.

To my father, whose endless patience with my endless questions, and constant emphasis on reaching one's full potential through honesty and hard work, guided my aspirations.

To my sister, whose hard-earned success in life unlocked closed doors for me, and exceptional problem-solving skills helped solve so many of mine.

To my nephew and niece, whose refreshing innocence and heartfelt warmth
never failed to brighten up my life.

Last but by no means the least, to my love, companion, and partner, whose guiding light chaperoned my every step, tirelessly uplifting my downward-trending spirits,
and unwavering faith in me persuaded me to have some faith too.

To all of them, for their steady love, support, and guidance that kept me pushing forward through all the hurdles along the way – this one is for you.

ACKNOWLEDGEMENT

By the Grace of God, I was fortunate enough to receive the help and support of a large number of individuals who made this dissertation possible. First of all, I would like to express my deepest gratitude to my research advisor and committee co-chair, Dr. Graciela Gonzalez, who was my main source of inspiration for embarking on, and successfully concluding, this PhD journey. Dr. Gonzalez provided me her expert research advice, while still granting me the freedom to explore new ideas of my own, and actively helped me acquire funding through research assistantship positions every year. She kept her faith in me, even when I had none left, and her steady support was paramount in helping me move forward in research and in life. I would also like to sincerely thank my research advisor and committee co-chair, Dr. Matthew Scotch, whose guidance was vital in driving my PhD. Dr. Scotch, too, supervised and supported my research initiatives, and hired me as a research assistant, year after year. He never failed to quickly address any issue I sought his help with and provided invaluable feedback that allowed me to learn and grow. I am also very grateful to my committee member, Dr. George Runger, for his continued encouragement, support, and advice that was key in helping me complete this dissertation.

I would also like to thank all the other co-authors of the published work included in this dissertation for their assistance with the publication process, as well as for their permission to use our work here. Among them, I would especially like to thank my husband, Dr. Robert Rivera; in addition to patiently supporting me in every way possible as my loving life-partner, Dr. Rivera also played a key role in structuring the framework of this research, significantly helped with annotation efforts, and read and commented on this entire dissertation. I am also very grateful to Dr. Davy Weissenbacher, whose assistance was of utmost importance in this work. Dr. Weissenbacher co-authored with me every research paper included here, guided me through different research initiatives, and sent me numerous dissertations that he thought might be relevant to my work to help me write mine when the time came. I would also like to sincerely thank Rachel Beard, Mari Firago, Rob Lauder, and Karen O'Connor for their annotation support, which played a significant role in my research, and Dr. Arjun Magge for his support near the end. I am also very grateful to Dr. Garrick Wallstrom, Dr. Daniel Magee, Demetri Jones-Shargani, and Matteo Vaiente for their assistance.

Numerous other faculty, staff, colleagues, and friends at ASU also assisted me in various ways throughout my PhD. I would like to especially thank Dr. Davide Sottara for his support which played an instrumental role in helping me continue my work. I would also like to thank Dr. Chitta Baral, Dr. William Johnson, Dr. Abeed Sarkar, Dr. Ehsan Emadzaddeh, Dr. Azadeh Nikfarjam, and Dr. Robert Leaman for their aid and guidance at different stages of my PhD journey. I am also truly grateful to Lauren Madjidi, Maria Hanlin, Laura Kaufman, Patricia Hutton, and Kaitlin Yacob for helping me keep track of the numerous deadlines related to the PhD program and schedule events accordingly. Special thanks to Lauren Madjidi for playing such a key role during the final stretch, and to Melanie Taussig and Michael Crandall for their critical support in the first year. I would also like to thank all my colleagues and friends while at ASU for their support esp. Dr. Sheaza Ahmed, Neel Mehta, Dr. Gazi Islam, Shah Enam, and Dr Prabal Khanal. My deepest gratitude is to Dr. Sheaza Ahmed, who went above and beyond to help me out during my difficult first year.

This acknowledgment section is certainly incomplete without thanking the rest of my family and friends who played such a vital role in helping me get here: my parents, for loving, nurturing, and motivating me and instilling in me my guiding values; my sister for being an amazing sister in every way – from helping me write my graduate school essay to actively supporting me during the final long stretch, she was there for me all through; my nephews and nieces for brightening up my life; all my aunts, uncles and cousins for their love, support, and prayers; my brother-in-law for taking such great care of our family and rooting for me all through; my mother-in-law for supporting Robert and me in endless ways and encouraging me always; all my friends, esp. Maisha for standing by my side, as my trusted confidante, from the start to the end, staunchly refusing to give up on me no matter what; Tanisha for being my steady source of advice and inspiration on countless occasions; Subehee for helping me out at a moment's notice whenever I needed her; Jackie for her endless supply of encouragement and optimism throughout my PhD; Nausheen for always caring and making the effort to stay in touch; and Sarah, Reme, Shaaz, Piyal, and Saamiya for their support on so many occasions.

I would also like to acknowledge support from NIH NLM grant no. R01LM012080 and NIH NIAID grant no. R01AI117011.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Information extraction (IE) is a rapidly growing domain within text mining which addresses the challenge of extracting structured information about entities, relations, or events from unstructured or semi-structured textual data. In recent years, IE methods have been used to help accelerate research in different biomedical domains such as clinical decision support, pharmacogenomics and adverse drug reaction (ADR) monitoring [1]–[9]. However, relatively few works in IE have explored supporting biomedical research that utilizes geospatial metadata representing the sampling location of taxa, which in case of infectious pathogens refer to the location of their infected host (LOIH). Some examples of such research areas include phylogeography [10], spatial epidemiology [11] and infectious disease surveillance [12]. Currently, researchers typically retrieve this information through manual review of nucleotide sequence databases, such as GenBank [13], and relevant biomedical publications. However, the process of manually extracting this information is slow and challenging and represent a major bottleneck for studies incorporating this information [14]. Therefore, an automated system for geo-referencing genetic sequences could help significantly accelerate this process, positively impacting a wide range of biomedical research.

This work explores the use of knowledge-driven heuristics for geo-referencing genetic sequences based on metadata available in GenBank and related full-text publications. It specifically focuses on supporting the domain of virus phylogeography, which analyzes the geographic distribution and genetic variation of viruses and has the potential to play a significant role in infectious disease surveillance, vaccine design and distribution, and viral epidemiology [15]. However, the methods, datasets, and software presented here would also benefit diverse research studies outside the domain of virus phylogeography that require the sampling sites of genetic sequences.

This chapter introduces the problem addressed through this thesis, discusses related work and main contributions, and outlines the thesis organization. Section 1.1 explains the background and motivation for the specific focus of this research – extracting, integrating and normalizing the LOIH of viruses based on information available in GenBank and related full-text articles to help support

virus phylogeography; section 1.2 discusses related work in this field; and section 1.3 summarizes the specific contributions of this dissertation.

1.1 Background and Motivation

Phylogeography is the study of the geographical distribution of genealogical lineages. It allows researchers to model the migration patterns and genetic variation of genetic sequences over time and has recently grown into a popular means of studying pathogens such as viruses [16]. As one of the major causes of infectious diseases across the world, viruses represent a potent threat to population health, and have resulted in several pandemic and epidemic diseases over the past few decades [17]. Some examples include the 2003 SARS epidemic, the 2009 H1N1 influenza A pandemic, and the 2014 Ebola outbreak in West Africa. In fact, it has been estimated that viruses with RNA genomes account for a third of all emerging and re-emerging infections [18], and factors such as rising population density and increased global travel are expected to further increase the frequency and severity of such disease outbreaks. Therefore, understanding the evolutionary dynamics and geographical transmission of viruses, through diverse methods of analysis, including virus phylogeography, is of critical importance. By allowing researchers to estimate the origins and drivers of viral diseases, virus phylogeography plays a key role in infectious disease surveillance and virus epidemiology with the potential to significantly impact public health outcomes.

To construct phylogeographic models of virus diffusion, researchers require genetic sequence data, date of collection, and location of infected host (LOIH) of each virus included in the analysis. Phylogeographic models may be discrete [19] or continuous [20]. In continuous phylogeographic approaches, the specific latitude and longitude coordinates of the LOIH of each virus is included in the model. For instance, Brunker *et al.* [21] collected rabies virus (RABV) sequences from rabid animals in the Serengeti district of Tanzania, along with their GPS location, and applied a continuous phylogeography framework to study the effect of landscape attributes on RABV diffusion within this district. In discrete phylogeographic approaches, the LOIH of each virus is represented as a discrete state in the model and its required level of granularity depends on the scope of the study. For instance, Raghwani *et al.* [22] applied a discrete phylogeographic framework incorporating district-level geographic data to analyze the distribution of dengue virus within Ho Chi

Minh City in Vietnam. Such an analysis would not have been feasible using only country-level geographic data. Although discrete phylogeography studies may be performed at different levels of geospatial specificity (country-level, state-level, county-level etc.), Magee *et al.* [23] demonstrated that the most precise sampling locations available should be used in virus phylogeographic models to enable accurate analysis of virus diffusion predictors [23].

To retrieve the LOIH of viruses at their desired level of granularity, phylogeography researchers often start with reviewing the geographic metadata in sequence databases such as GenBank. GenBank is a popular database, developed and maintained by the National Center for Biotechnology Information (NCBI), for the submission and analysis of nucleotide sequences. It is part of the International Nucleotide Sequence Database Collaboration (INSDC), and includes genetic sequences submitted by researchers worldwide [24]. At the time of writing, GenBank contains over two million nucleotide sequences of virus origin along with predefined metadata containing information about each sequence. This includes geographic metadata denoting the location of infected host (LOIH) of the virus.

The designated field for storing the LOIH of a virus within a GenBank entry is called the *country* field which, despite its name, does not simply contain the name of the country from which the genetic sequence was isolated. Depending on the level of detail the submitter of the sequence chose to include, it may contain locations with varying degrees of granularity, or may be blank and contain no geographic information. For instance, in GenBank record with accession number *AB520868* [25] this field contains city-level geographic information: “Japan: Miyagi, Sendai”; in *CY000071* [26], it contains county-level geographic information: “USA: Greene County, NY”; in *AB518493* [27] it contains the country name “Brazil”; in *M63769* [28] it only contains the place name, “Cambridge”, which is highly ambiguous and can refer to one of many locations across the world; and in *M63757* [29] it is blank, containing no geographic information. Due to the specific nature of virus nomenclature, additional geographic metadata may sometimes be found in other fields of the record such as the *strain* and *isolate* fields. For instance, the *strain* field of GenBank record *JQ714202* [30] contains “A/Tianjinheping/SWL313/2009” while the *country* field contains “China,” thereby implying that the sequence was isolated from the Heping District within the Tianjin

province of China. Therefore, to retrieve adequate information about the LOIH of a virus from GenBank, researchers often need to integrate geographic metadata from multiple fields in the record. This can be a time-consuming and challenging process, especially when a researcher is not very familiar with the geographic region in which the study is being conducted.

Depending on the type and scope of phylogeographic analysis being performed, a researcher may require more specific geographic information about the LOIH of a virus than what is available in GenBank. For instance, it has been estimated that over 80% of GenBank records pertaining to RNA viruses within tetrapod hosts do not have geographic metadata more specific than the Administrative Division 1-level (ADM1), which is the state or province level [14]. Therefore, researchers developing precise phylogeographic models would consider the geographic metadata in the vast majority of GenBank records to be insufficient.

When the geographic metadata in GenBank is found to be insufficient for an analysis, researchers often search full-text publications linked to the records for more specific information. If available, the more specific metadata from the related article is incorporated within their analysis. However, a manual survey of articles is a time-consuming and tedious process and presents a major bottleneck for data collection. Moreover, since each publication typically represents a study related to multiple GenBank records, it may only include a generic sentence listing the different locations where the study was performed, without including a one-to-one-mapping between the locations and the related records. In such cases, it may be beneficial to test multiple models using the different possible LOIHs of each virus.

After retrieving the LOIH of all viruses being analyzed, phylogeography researchers need to normalize the LOIH of each virus based on the requirements of their specific model. Continuous phylogeographic models require the specific latitude/longitude coordinates of the LOIH of all viruses included in the model. Therefore, when developing such models, researchers need to manually map the LOIH of each virus to its correct geospatial coordinates using an online resource such as GeoNames [31], which lists over 10 million geospatial locations across the world. This is a highly challenging task, since some locations can be very ambiguous and possibly map to many different geospatial coordinates. For instance, according to GeoNames, “Bristol, USA,” can refer to a town

or county in one of over 15 states in the USA [32], and therefore, choosing the correct "Bristol" for GenBank record *KX029319* [33] which simply contains "USA: Bristol" in its *country* field, with no additional geographic metadata in the other record fields, is not easy. Different researchers may choose to resolve the ambiguities in different ways, and this may lead to the use of different geospatial coordinates for the same record in different studies.

Discrete phylogeographic models require each unique location to be represented as a discrete state in the model. Therefore, the string representation of the LOIH of each virus in the discrete model must be normalized by creating identical string representations for identical locations, and non-identical string representations for mutually exclusive locations. For instance, the US state of New York cannot be represented by "New York" in one case and "New York State" in another. This can be a problem when different spelling variants are used to represent the same location in different sources. For instance, in an analysis involving 706 GenBank records, we found over five different spelling variations of the Egyptian governorate of "Beni Suef" in GenBank and they each needed to be normalized to the same string representation. Another possible problem may arise if a model includes two different places with the same name. For instance, "Paris" may denote the capital city of France or a city in Texas, USA, and if both of these locations are included in a model, it is important to ensure that two different string representations are used for the two locations.

Given the wide range of challenges involved in manually extracting, integrating, and normalizing the LOIH of viruses from GenBank record metadata and related full-text articles, an automated framework for performing these tasks could significantly accelerate this process and make it more systematic. The work described here represents the most comprehensive effort to address the construction of such a framework and has the potential to have a notable impact on public health research. Our study led to the development of a freely accessible system called GeoBoost for georeferencing virus GenBank records, as well as a publicly available database containing over two million virus GenBank records that were georeferenced using the methods introduced here.

1.2 Related Work

A considerable volume of work has been published in related research areas of genetic sequence annotation, bacteria location extraction, and toponym resolution (detection and disambiguation) in full-text articles related to GenBank. However, as we described, none have developed an end-to-end pipeline for geospatial metadata enrichment of viruses as we propose in this work.

Genetic sequence annotation is an active research problem which has been the focus of many research efforts in recent years [36]–[42]. Researchers have attempted to enhance the quality of genetic sequence metadata, such as the date, host and site of collection of taxa, in existing sequence repositories using both manual and automated curation efforts. This has led to the development of several structured databases and bioinformatics pipelines such as PATRIC [40], Virus Variation Resource [41], Virus Pathogen Resource [42], and SeqenV [43]. However, few works have focused specifically on normalizing the geographic metadata of GenBank records. The research studies included here represent the only efforts to apply an automated approach for geo-referencing all virus-related GenBank records.

The Bacteria Biotope task of the BioNLP Shared Tasks 2011, 2013 and 2016 [44]–[46] introduced the challenge of extracting events between bacteria entities and their locations (either habitat or geographical entities) from scientific web pages and PubMed abstracts, providing a fairly large manually annotated corpus for this task. Researchers have applied different machine learning models, including state-of-the-art neural network models, to address this challenge. However, similar work has not been reported for viruses using full-text PMC articles.

Recently, the research problem of toponym resolution (detection and disambiguation) in full-text articles related to GenBank records has attracted a lot of attention, especially following the SemEval-2019 Shared Task 12 [47]. Many of the teams participating in this task achieved a high f-score for toponym detection in a corpus of GenBank-related full-text articles with the top performing system [48] achieving a strict micro-averaged f-score of 0.9161. Also, Magge *et al.* [49] recently applied a bi-directional neural network model for toponym detection in a subset of the corpus provided for this task and achieved an f-score of 0.94. This represents a significant

improvement of 0.24 over the f-score of the primarily dictionary-based NER system used in GeoBoost for toponym detection, as described by Weissenbacher *et al*. [50]. However, since GeoBoost applies various knowledge-based constraints prior to linking geographic locations extracted from full-text articles to their corresponding GenBank records, to account for the low precision of its NER component, it is unclear whether this performance improvement in toponym detection would also lead to a notable improvement in GeoBoost's performance.

Magge *et al.* also achieved a slightly higher disambiguation accuracy (91%) than what was reported for GeoBoost (88%) in full-text articles by adopting the approach taken by Tamames *et al.* [51] of including information about parent locations, when present in contiguous text following a place name, in addition to incorporating the features used by GeoBoost's toponym disambiguation algorithm. However, this might not have a significant impact on GeoBoost's accuracy given the way the system is currently designed. Since GeoBoost extracts and integrates toponyms linked to GenBank records from different information sources, it only disambiguates the final integrated geographic metadata after all extraction and integration steps have been performed to maximize accuracy. Therefore, although GeoBoost provides users the option of using its NER system to disambiguate the names of places in full-text articles, it does not utilize the disambiguation results to complete its primary objective of normalizing the LOIH of viruses in GenBank and it is not clear whether incorporating lexical context from the paper would help boost its disambiguation accuracy. GeoBoost's disambiguation techniques are based on established knowledge-based methods of named entity linking [52]–[54] but, to the best of my knowledge, other studies have not evaluated these methods within this specific domain.

1.3 Specific Contributions

This dissertation is a compilation of the following three published works describing knowledge-driven methods for geographic information extraction in the biomedical domain which we developed to address the primary research objective of geo-referencing GenBank records for supporting virus phylogeography:

- Tahsin, T., Weissenbacher, D., Rivera, R., Beard, R., Firago, M., Wallstrom, G., Scotch, M. and Gonzalez, G., 2016. A high-precision rule-based extraction system for expanding geospatial metadata in GenBank records. *Journal of the American Medical Informatics Association*, 23(5), pp.934-941

  This work, which constitutes Chapter 2 of this dissertation, describes and evaluates the knowledge-driven heuristics we used to develop the basic infrastructure of GeoBoost, a system for extracting, integrating, and normalizing the geographic metadata of virus GenBank records based on information present in the records and related full-text articles. My specific contributions in this publication include composing the initial draft of the article, revising its content and performing additional experiments based on reviewer feedback, and developing and evaluating all components of the geographic information extraction system described there aside from the Text Parser, which was responsible for toponym detection in full-text articles linked to GenBank. The components I developed and evaluated include: 1) Record Location Extractor which extracted and integrated geographic metadata from the "country", "strain" and "isolate" fields of GenBank records, 2) Sufficiency Analyzer which detected whether the existing geographic metadata in a GenBank record satisfied a pre-defined sufficiency criteria of geographic granularity, 3) Table Linker which linked names of places in the tabular content of relevant full-text articles to their respective GenBank records, 4) Text Linker which linked toponyms extracted by the Text Parser to their corresponding GenBank records, 5) Data Integrator which integrated all the geographic locations linked to each GenBank record by the Record Location Extractor, Table Linker and Text Linker to output the most probable LOIH of each virus based on knowledge-driven heuristics, and 6) Location Disambiguation Module which mapped the integrated geographic metadata produced by Data Integrator to its corresponding latitude/longitude coordinates.

- Tahsin, T., Weissenbacher, D., Jones-Shargani, D., Magee, D., Vaiente, M., Gonzalez, G. and Scotch, M., 2017. Named entity linking of geospatial and host metadata in GenBank for advancing biomedical research. *Database*, 2017.

This work, which constitutes Chapter 3 of this dissertation, describes the development and evaluation of a publicly available database containing the normalized geographic and host metadata of all virus-related GenBank records downloaded at the time of publication by our research team. We normalized the geographic metadata for the GenBank records using an updated version of the Record Location Extractor introduced in Chapter 2 and the research methods used for this purpose form a central component of this thesis. My specific contributions in this publication include composing the initial draft of the article, assisting with content revision and additional experiment completion based on reviewer feedback, developing and applying the pipeline for normalizing the geographic metadata of GenBank records to help build the database presented in the paper, calculating pertinent database statistics related to geographic metadata normalization in the database, calculating the annotation statistics for the manually-annotated test set, and evaluating the normalized host and geographic metadata in the database using this test set.

- Tahsin, T., Weissenbacher, D., O'connor, K., Magge, A., Scotch, M. and Gonzalez-Hernandez, G., 2017. GeoBoost: accelerating research involving the geospatial metadata of virus GenBank records. *Bioinformatics*, 34(9), pp.1606-1608.

  This work, which constitute Chapter 5 of this dissertation, introduces GeoBoost, a publicly available desktop application for automatically extracting, integrating, and normalizing the LOIH of viruses based on information present in GenBank records and related full-text articles, and assigning confidences estimates to each possible LOIH of the virus. It describes the knowledge-driven heuristics used by GeoBoost to complete the research objectives of this dissertation and presents its performance on a manually annotated test set. My specific contributions in this publication include composing the initial draft of the article, performing content revision based on reviewer feedback, and developing and evaluating the entire GeoBoost framework presented there. It includes enhanced versions of the modules described in Chapter 2, including a more efficient implementation of the *Text Parser*, and incorporates additional modules to improve the pipeline. The different modules I developed in GeoBoost can be divided into three layers: 1) Data Acquisition

Layer which is responsible for downloading relevant data, 2) Knowledge Layer which is responsible for maintaining a Lucene index of geographic locations containing knowledge derived from the GeoNames database and performing spatial reasoning based on knowledge-driven heuristics, and 3) Logic Layer which is responsible for utilizing the geographic knowledge provided by the Knowledge Layer, and the data downloaded by the Data Acquisition layer to output: i) the most probable, integrated, normalized location of infected host (LOIH) of each virus and ii) the probability scores of each possible LOIH of each virus (confidence estimate output).

The three publications listed above progressively advance the central objective of this dissertation – researching knowledge-driven geographic information extraction methods for geo-referencing virus GenBank records based on information present in the record and related full-text articles. In addition to these publications, I have also contributed significantly to additional published work relevant to this thesis. In [55], I evaluated state-of-the-art NER tools for extracting species, gene, and temporal mentions from full-text articles linked to virus GenBank records and measured their performance using annotations performed by our research team. Additionally, since existing NER tools for extracting geographic mentions from text were previously found to perform poorly in this domain, I also developed and tested a new, efficient, lexicon-based approach with promising results. In [50], I developed the "metadata heuristic", a knowledge-based method for disambiguating names of places in full-text articles linked to GenBank records which was found to notably increase the performance of existing toponym disambiguation heuristics when evaluated within this specific biomedical domain, and estimated that this heuristic can potentially impact toponym disambiguation in over 200,000 PubMed articles. A few other works I co-authored which are related to the research objectives of this dissertation include [56] and [34].

2 A HIGH-PRECISION RULE-BASED EXTRACTION SYSTEM FOR EXPANDING GEOSPATIAL METADATA IN GENBANK RECORDS

Authors: Tasnia Tahsin, Davy Weissenbacher, Robert Rivera, Rachel Beard, Mari Firago, Garrick Wallstrom, Matthew Scotch, and Graciela Gonzalez

2.1 Abstract

Objective: The metadata reflecting the location of the infected host (LOIH) of virus sequences in GenBank often lacks specificity. This work seeks to enhance this metadata by extracting more specific geographic information from related full-text articles and mapping them to their latitude/longitudes using knowledge derived from external geographical databases.

Materials and Methods: We developed a rule-based information extraction framework for linking GenBank records to the latitude/longitudes of the LOIH of viruses. Our system first extracts existing geospatial metadata from GenBank records and attempts to improve it by seeking additional, relevant geographic information from text and tables in related full-text PubMed Central articles. The final extracted locations linked to the records, based on data assimilated from these sources, are then disambiguated and mapped to their respective geo-coordinates. We evaluated our approach on a manually annotated dataset comprising of 5728 GenBank records for the influenza A virus.

Results: We found the precision, recall, and f-measure of our system for linking GenBank records to the latitude/longitudes of their LOIH to be 0.832, 0.967, and 0.894, respectively.

Discussion: Our system had a high level of accuracy for linking GenBank records to the geo-coordinates of the LOIH of viruses. However, it can be further improved by expanding our database of geospatial data, incorporating spell correction, and enhancing the rules used for extraction.

Conclusion: Our system performs reasonably well for linking GenBank records for the influenza A virus to the geo-coordinates of their LOIH based on record metadata and information extracted from related full-text articles.

2.2 Background and Significance

Information extraction (IE) involves the use of natural language processing techniques for automated extraction of structured information about entities, relations, or events from unstructured textual data. In recent years, the rapidly expanding field of IE has been applied to accelerate research in various biomedical domains. For instance, IE methods are currently being used to automatically extract relations between drugs, genes, and diseases from PubMed articles in order to populate the structured PharmGKB database [1], which in turn can be used for advancing personalized medicine.

Much less work in IE has explored supporting public health applications that heavily rely on detailed geospatial information about the sampling sites of genetic sequences. One example of such an application is phylogeography, which has recently grown into a popular means of tracking the spread of infectious pathogens and enhancing their epidemiological analysis [2]– [4]. For instance, Hovmöller *et al* [5] used multiple phylogenetic trees to estimate the geographical transmission routes of a highly pathogenic strain of H5N1 virus and developed a web application for visualizing the estimated routes. Similarly, Janies *et al.* [6] combined phylogenetic analyses with visualization techniques to study the global spread of H7 influenza A viruses. In addition to advancing the surveillance of infectious diseases, such forms of sequence-based analysis, incorporating the location of the infected host (LOIH) from which the pathogen sequence was isolated, can also assist the design and distribution of vaccines, and help clinical researchers better understand the etiology of various diseases [2], [7], [8].

The geographic metadata required for studies involving the spatial modeling of sequences are often obtained from public databases such as GenBank [9], which is part of the International Nucleotide Sequence Database Collaboration and includes data deposited by researchers all over the world. Each GenBank record contains separate fields for holding various forms of sequence-related metadata such as strain name, date of collection, LOIH, and the type of host. The LOIH is typically present in the *country* field of the record. Despite its name, this field does not simply contain the name of the country in which the host was found; it may include locations with varying levels of specificity or may not contain any location at all. For instance, in GenBank

record *AY282759* (of note, this is the accession number of the record, not the GenBank identifier), this field does not contain any geographic metadata; in *M63769* it contains a single ambiguous place name, “Cambridge” without indicating the specific country in which it resides; in *GU332632* it contains the name of a state “USA:Iowa”; and in *CY024354* it contains the name of a city “China:Shantou.” Since building precise spatial models often requires very specific information about the LOIH of the sequences being studied, the geospatial metadata in GenBank, even when available, may not be sufficient for the researcher. For instance, Raghwani *et al*. [10] used district-level phylogeographic analysis to study dengue virus migration within Ho Chi Minh City in Vietnam. With only country-level or province-level geospatial metadata, such an analysis would not have been feasible.

Because of the absence or inadequacy of geospatial metadata in GenBank records, researchers might need to search full-text publications linked to the records for more specific information. If found, the more specific metadata from the paper is incorporated within their study. However, a manual survey of articles is a time-consuming and tedious process and presents a major bottleneck for data collection. Moreover, many studies may require the specific latitude/longitude coordinates of the sampling sites; thus, simply finding the name of the locations may not be enough. For instance, for continuous phylogeography studies [11] and disease spread visualization tools [12], obtaining the specific geo-coordinates is crucial. In this case, researchers wishing to use the information would need to perform an additional step of mapping each location to its correct geospatial coordinates using a database such as GeoNames [13], which lists 10 million geospatial locations across the world. This is not a trivial process, since some locations can be highly ambiguous and possibly map to a large number of unique coordinates. Consider, for example, “Malang, Indonesia,” which is mapped in GeoNames to 23 distinct locations. An additional problem ancillary to the manual process is that, depending on how the ambiguities are resolved, it is possible that different coordinates would be derived for the same study by different researchers.

Therefore, an automated system for the extraction of geospatial metadata from GenBank records and related full-text articles can help make this process faster and more systematic, positively impacting public health research. To the best of our knowledge, no such system currently

exists. The *Bacteria Biotope* Task of BioNLP Shared Task 2013 [14] included the extraction of localization relations between bacterial species and geographical entities, but participants were not required to map the geo-entities to their latitude/longitude coordinates and the task corpus consisted of web documents rather than full-text scientific articles. Additionally, Tamames and Lorenzo [15] performed toponym (location name) resolution (detection and disambiguation) in full-text articles mentioning the sampling sites of bacterial sequences, and achieved a precision and recall of 0.92 and 0.86, respectively, for this task. However, their work did not focus on linking the sequences to their collection sites in the articles. Other studies have analyzed or used GenBank metadata, often in combination with other resources, for various applications [16]– [20] but none specifically attempted the enhancement of geospatial metadata in GenBank using information extracted from full-text articles.

2.3 Objective

The objective of the present study was to provide an IE framework for automatically linking GenBank record sequences to the latitude and longitude coordinates of their LOIH in order to help advance public health research. Our system attempted to make this location as specific as possible by extracting geospatial metadata from GenBank record fields and related full-text PubMed Central (PMC) articles. As our primary case study, we present a detailed evaluation of our system on a set of manually annotated GenBank records for the influenza A virus. We chose this virus because of the large sample of influenza A sequences in GenBank as well as its significance in public health research. In addition, to test the generalizability of our system, we also report its accuracy on a smaller sample of GenBank records for St. Louis Encephalitis (SLE), Eastern equine encephalitis (EEE), Western equine encephalitis (WEE), West Nile virus (WNV), rabies, and hantavirus, which are some of the most widely studied zoonotic viruses (viruses transmittable between animals and humans) by public health, agricultural, and wildlife state departments in United States [21].

2.4 Materials and Methods

Our methodology for conducting this study can be divided into three broad stages: selection and download of GenBank records and related PMC articles, development of the IE system, and evaluation of the IE system. Each of these stages are described in detail below.

2.4.1 Selection and Download of GenBank Records and Related PMC Articles

For the influenza case study, we used stratified random sampling to select ~10% of all PMC articles linked to GenBank records for the influenza A virus (stratification was performed based on the number of records linked to each article; further details given in Appendix A). This produced a corpus of 60 PMC articles corresponding to 5728 GenBank records. We manually downloaded the PDF versions of these papers and used Xpdf [22] to convert them into text files (of note, we chose to use the PDF version of each article instead of parsing its HTML version or searching for its XML version in PMC Open Access because not all articles have an HTML version and only a limited subset of PMC articles are available through PMC Open Access). In addition, we used the National Center for Biotechnology Information (NCBI) Entrez Utilities application programming interface (API) [23] to download relevant metadata fields from the selected GenBank records, including: *country* , *strain* , *organism* , *isolate, date*, and *host* .

For our secondary study, we first gathered a list of PMC articles linked to at least 10 GenBank records associated with the remaining six viruses (SLE, EEE, WEE, WNV, rabies, and hantavirus) and randomly selected two PMC articles for each virus. For each selected article, we randomly selected 10 records for inclusion within our evaluation sample. This resulted in a total of 120 records from 12 articles, equally distributed among the 6 viruses.

2.4.2 IE System Development

Our geospatial information extraction (IE) system is largely dependent on the GeoNames [13] database, a large collection of over 10 million geospatial locations across the world which has been effectively used in several existing systems for toponym resolution [15], [24], [25]. In addition to location names, GeoNames also contains several useful features about each entry in the database such as population data, country code and the latitude and longitude coordinates of the location's

centroid. For the purpose of this project, we downloaded the GeoNames data available online and imported it into a local database. In addition, we also imported data from the Socrata dataset [26], which contains geospatial data for 243 countries, since several country names were too ambiguous in GeoNames. For instance, the results from the query "Italy" in GeoNames does not include the country "Italy" – it contains populated places in other countries. One needs to query for "Repubblica Italiana" to retrieve this country. France, on the other hand, is listed as "Republic of France" and not "Republique francaise". If we include alternate names for these countries we obtain names such as "Farani" for France, which are not as likely to be referring to the country in a scientific article written in the English language and may generate more false positives. Therefore, we opted to include the Socrata dataset which focuses solely on countries.

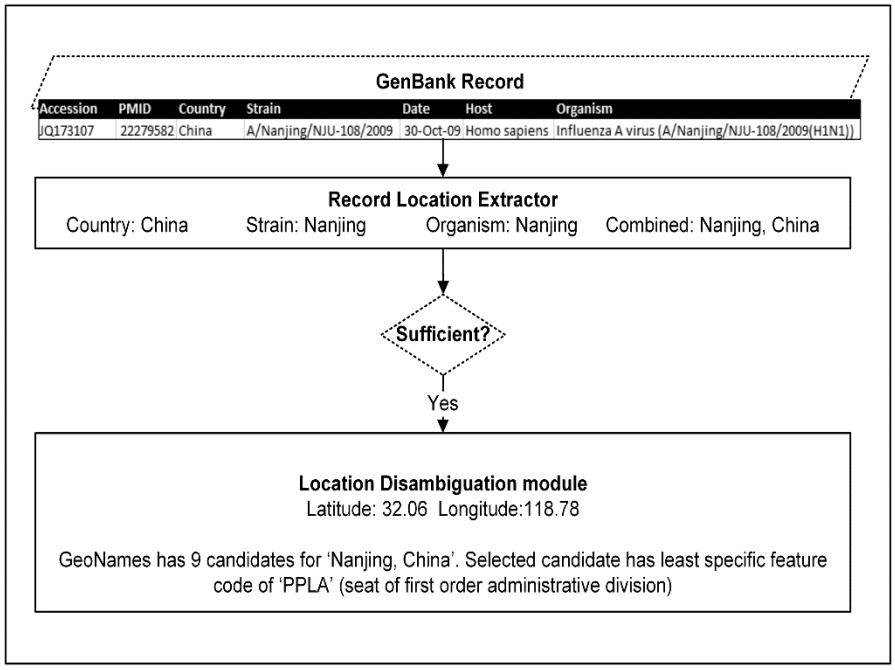


Figure 1. System pipeline for a GenBank record with sufficient geospatial metadata (location of infected host more specific than state or province level).

In order to introduce a stopping criterion for our system, we defined any location more specific than ADM1 (first order administrative division) level as "sufficient" based on our prior study [27]. This includes counties, districts, cities, towns or any other form of populated place within a country which is less specific than states and provinces. As illustrated in Figure 1, if a record already

contains sufficient geographic metadata within the record, our system directly proceeds to assigning geo-coordinates to the locations present in the record metadata instead of processing the related paper. For records with insufficient geographic metadata, it attempts to extract more specific information from the related article, if available, until it finds a LOIH that is considered sufficient (see Figure 2). Therefore, our system is capable of finding locations more specific than ADM1 (such as districts and cities) but does not necessarily search for more specific geospatial information once it finds such a location. To ensure retrieval of even more specific locations, when available, the stopping criterion would need to be altered.

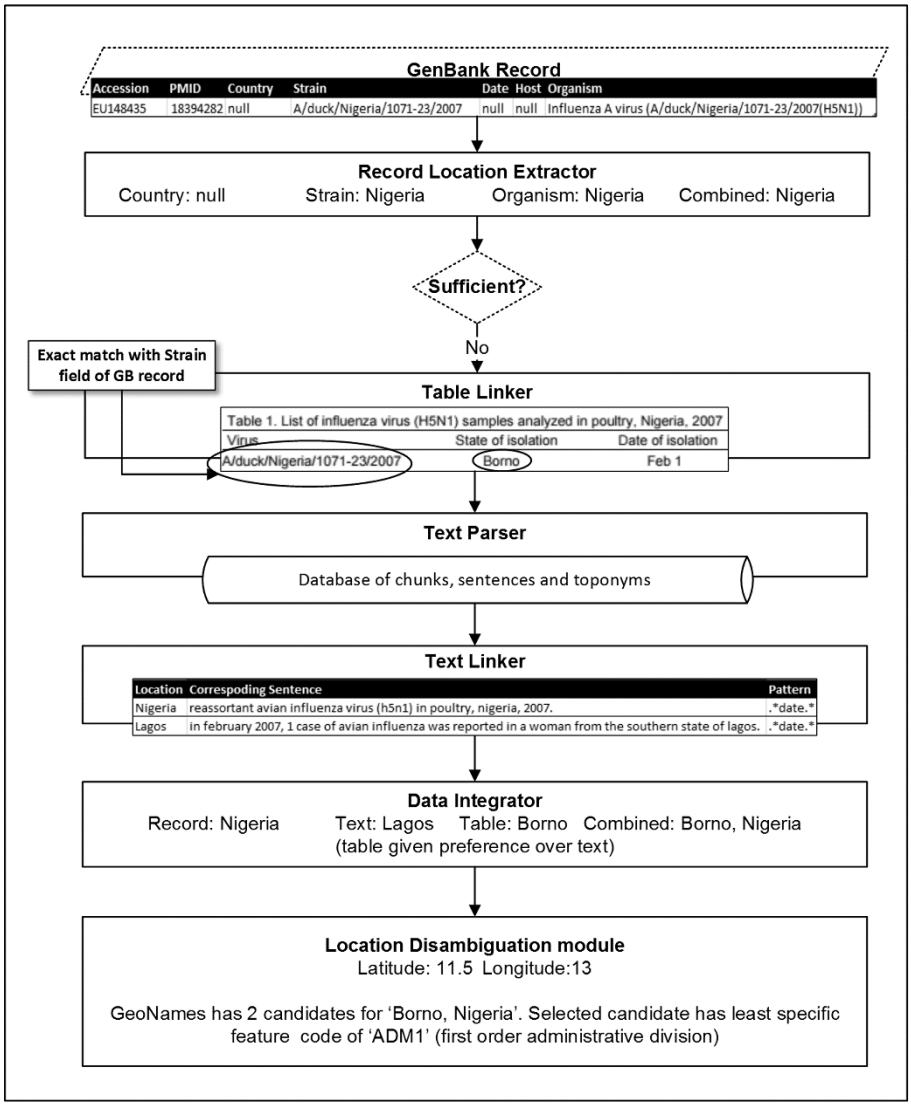


Figure 2: System pipeline for a GenBank record with insufficient geospatial metadata (location of infected host not more specific than state or province level).

For our current study, we integrated data from the *strain, isolate* and *organism* fields of GenBank records, in addition to the *country* field, for extracting existing geospatial metadata from the records. This is because virus nomenclature often includes the unique practice of incorporating the LOIH of sequences within their taxon names. For instance, the strain field of GenBank record JF340084 contains "A/St.Petersburg/14/2010" while the country field contains "Russia", thereby implying that the sequence was isolated from "St. Petersburg, Russia". The individual steps in the system pipeline are described below.

2.4.2.1 Record Location Extractor

The first step in our system pipeline involves the integration of existing geospatial metadata present within the *country*, *strain*, *isolate*, and *organism* fields of the GenBank record in order to identify the most specific LOIH available within the record itself. For instance, if the *country* field contains the location "China" while the *strain* field contains the location "Guangdong," then the most specific LOIH for this record based on record data will be "Guangdong, China." The record location extractor module uses our integrated database of geospatial locations for detecting location names mentioned within the record, determining the administrative level of each location detected and pairing extracted location names with any mention of their parent ADM1-level location and/or parent country in the record.

2.4.2.2 Sufficiency Analyzer

Depending on the final output produced by the record location extractor, a GenBank record may be classified as either *sufficient* or *insufficient*. To be considered sufficient, the GenBank record must either contain a location name more specific than ADM1-level along with the name of the country in which it resides; e.g., "San Diego, USA" or a location name more specific than ADM1-level that can only be present in a single country; e.g., "Beijing," which can only be present in "China." A record classified as *insufficient* may contain no location information, only country-level information; e.g., "China," ADM1-level information along with associated country name; e.g., "Guangdong, China" or ambiguous location information for which no matching country name could be found within the record; e.g., "Osaka," which is a place that can be present in Japan, USA, South

Africa, or the Solomon Islands. For all records that are classified as *insufficient*, our system searches the related full-text PMC article for more specific geospatial information using our table linker, text parser, and text linker modules.

2.4.2.3 Table Linker

The purpose of the table linker module is to identify possible LOIHs using table data (if available) in referenced PMC articles. Since the conversion from PDF to text does not allow the tables in the document to retain their defined structure, we directly parsed the HTML content of the articles in PMC to extract table information. For every table in each article related to at least one insufficient record, the table linker analyzes the table headers to determine whether or not it possesses relevant information that could be used to link a GenBank record to a geographic location. A table is considered relevant if one of its headers contains the word *location* or any of its synonyms *(location column)* and another contains the words *accession*, *strain*, *date*, *host*, or one of their synonyms *(GenBank metadata column)*. We manually compiled a list of synonyms for each of these words. If the data from the *GenBank metadata column* of one of the rows of a relevant table matches the metadata of the record, the table linker links data from the *location column* of the row to that record (see Figure 2). Aside from date metadata, we used exact match as our means of establishing equivalency between table data and GenBank record data. For dates, we first normalized the data before comparing them. Here, we used Stanford SUTime [28] along with a separate rule-based program that we wrote for normalizing date metadata in GenBank records, presented in formats such as “12-May” and “12-Jun-2007,” which Stanford SUTime was unable to parse, into TIMEX expressions.

2.4.2.4 Text Parser

To allow effective IE from the textual content of the article, we first used the text parser module to extract all sentences, tokens, and toponyms and stored them in a local database. For toponym detection, we used our system presented by Weissenbacher *et al*. [29], which was found to have a precision, recall, and f-score of 0.599, 0.904, and 0.72, respectively, for this task. For sentence

segmentation, we used the ANNIE module from the GATE platform [30], while for word segmentation, parts-of-speech tagging, and chunking, we used the Genia tagger [31].

2.4.2.5 Text Linker

The text linker module links the geospatial entities identified by our toponym detection system within an article to relevant GenBank records using a rule-based approach. A geospatial location extracted from an article is linked to a record referencing the article if the sentence containing the location also mentions other record metadata, such as strain name and accession number, or fits a few simple patterns that we developed. In Appendix B, we list the patterns used by our system. For every geospatial entity identified by our toponym detection system in an article related to a specific insufficient record, our text linker first determines whether the entity is present in a relevant section of the article (of note, the *Author Affiliation*, *Acknowledgment*, and *Reference* sections are considered to be the only "irrelevant" sections of an article) and for those that are, it proceeds to analyze the sentence containing the entity to check if it fits any of the utilized patterns; if it does, the entity is considered to be a possible candidate for the location of the virus.

2.4.2.6 Data Integrator

The data integrator module assimilates information extracted by the record location extractor, table linker, and text linker modules for a given GenBank record to produce a coherent set of geographical locations that are possible LOIHs for the record. This set of locations is referred to as final locations in the remainder of the article. The first step in this process involves the elimination of all locations extracted by the text linker and table linker which are inconsistent with the output produced by the record location extractor. A location is said to be inconsistent if one of the following is true: 1) It does not belong to the same country as the record location 2) It does not belong to the same ADM1-level location as the record location (*e.g.* "Philadelphia" is inconsistent with "Arizona, USA"). Once all the inconsistent locations have been removed, the data integrator uses output from the table linker and text linker to increase the specificity of the record location until a sufficient location is found; if no sufficient location is found, it outputs the most specific location retrieved. In case of locations with the same level of specificity, preference is given to table-derived locations

over text-derived locations. This is because tables tend to link each individual record to its precise LOIH (one-to-one mapping) whereas paragraphs in the article typically provide a list of locations related to all records referenced in them (see Figure 2). Therefore, adding information from the text linker, when sufficient table data is present, may reduce system precision. The final output from this module consists of distinct, non-overlapping locations considered by the system to be the set of most specific LOIHs available for the record based on the sources analyzed. This may include more than one location if the heuristics fail to narrow it down to a single LOIH. For every location, the parent country name and ADM1-level location is also included in the output, if found by the system.

2.4.2.7 Location Disambiguation Module

The location disambiguation module links each final location to its specific latitude and longitude coordinates. First, it queries our geospatial database to retrieve all possible latitude/longitudes for the location. Next, it sorts them based on their feature codes (code in GeoNames denoting the type of the location e.g. country, state, city etc.*)* and chooses the group of coordinates belonging to the least specific feature codes. For instance, in GeoNames, "Arizona" can be both a state in USA with feature code of *ADM1* and a populated place in the state of Texas, USA with feature code of *PPL* but our system will only select the former since it has a less specific feature code. This heuristic is based on the assumption that in the majority of cases, when an author mentions a location name which can refer to multiple places on earth with varying levels of specificity, he/she is referring to the one which is more widely known across the world and the less specific a place is, the more widely known it tends to be. Lastly, the module sorts the group of coordinates selected in the previous step based on their population, and outputs the set with the highest population. This is a popular heuristic [32] within the field of toponym disambiguation since it is generally assumed that places which have higher populations are better known among people and are more likely to be mentioned. When querying the database for the latitude/longitudes of a given location, we included the country code and ADM1 code of the location, if known.

The location disambiguation module links each final location to its specific latitude and longitude coordinates. First, the system queries our geospatial database to retrieve all possible latitude/longitudes for the location. Next, it sorts them based on their feature codes (code in

GeoNames denoting the type of the location, e.g., country, state, city, etc.) and chooses the group of coordinates belonging to the least specific feature codes. For instance, in GeoNames, "Arizona" can be both a state in United States with feature code of *ADM1* and a populated place in the state of Texas, United States with feature code of *PPL* but our system will only select the former since it has a less specific feature code. This heuristic is based on the assumption that in the majority of cases, when an author mentions a location name that can refer to multiple places on earth with varying levels of specificity, he/she is referring to the one that is more widely known across the world and the less specific a place is, the more widely known it tends to be. Lastly, the module sorts the group of coordinates selected in the previous step based on their population, and outputs the set with the highest population. This is a popular heuristic [32] within the field of toponym disambiguation since it is generally assumed that places that have higher populations are better known among people and are more likely to be mentioned. When querying the database for the latitude/longitudes of a given location, we included the country code and ADM1 code of the location, if known.

2.4.3 System Evaluation

In order to evaluate our system for the influenza case study, three annotators manually annotated the 5728 GenBank records linked to the 60 related papers. The annotators followed a set of guidelines created prior to development of the corpus and documented relevant data for evaluating each individual module within the system pipeline (see Appendix C for annotation details and Figure 3 for example). We calculated the inter-rater agreement (IRR) for "sufficiency" annotation of 2017 records related to six randomly selected PMC articles and the final location (defined in the *Data Integrator* section) annotation of 1477 records related to 36 randomly selected PMC articles. We used the traditional IRR metric of Cohen's kappa statistic as our measure of IRR for "sufficiency" annotation. However, for the other annotations, we used f-score as our measure of IRR, holding one annotator as the gold standard each time. This is because Cohen's kappa calculation requires well-defined negative cases, which we lack for these annotations, and f-score has been shown to be a reliable IRR measure for information retrieval tasks [33]. After calculation of the IRR, the annotators performed multiple rounds of annotation for all records to ensure that the

guidelines were followed, and any mistakes corrected before creation of the gold-standard corpus (included as supplementary file in Appendix D).

Accession No.	PMID	Sufficient?	Record Locations	Text Locations	Table Locations	Final Locations	Latitude	Longitude
DQ073419	16306617	No	Henan; China	Pingyu; Henan; China		Pingyu, Henan, China	32.999	114.611
EU148363	18394282	No	Nigeria	Plateau; Nigeria	Plateau	Plateau, Nigeria	9.167	9.75
EU148364	18394282	No	Nigeria	Sokoto; Nigeria	Sokoto	Sokoto, Nigeria	13.083	5.25
EU148372	18394282	No	Nigeria	Borno; Nigeria	Borno	Borno, Nigeria	11.5	13
FJ461592	19359528	No	South Korea	Chungnam; South Korea; Korea		Chungnam, South Korea	36.5	127
EU544242	18704172	No	Nigeria	Hadejia-Nguru Wetlands; Jigawa; Nigeria		Hadejia-Nguru Wetlands, Nigeria; Jigawa, Nigeria	12.48	10.44
GU201599	21645421	No	India	Gwalior; Madhya Pradesh; India		Gwalior, Madhya Pradesh, India	26.229	78.173
HM114446	20592108	No	Vietnam	Vietnam	Ha Nam	Ha Nam , Vietnam	20.533	105.966
HM114526	20592108	No	Vietnam	Vietnam	Thai Binh	Thai Binh , Vietnam	20.45	106.34
HM114542	20592108	No	Vietnam	Vietnam	Vinh Phuc	Vinh Phuc , Vietnam	21.333	105.566

Figure 3. Example of annotated GenBank records for the influenza case study.

We evaluated the different components of our system for the influenza case study using Exact Match per record-location linkage criteria. In this case, a true positive indicates that a given record-location linkage extracted by our system is equivalent to one annotated in the gold standard. For the final locations, we normalized the country codes to allow for fairer comparison but partial matches were not allowed. For instance if the annotated final location for a record is “New York City, New York, USA” and our program simply outputs “New York, USA,” then we count this linkage as both a FP and a false negative. When evaluating the disambiguation module, we considered two geo-coordinates of non-country locations to be equivalent if they were within 10 miles of each other; geo-coordinates of country-level locations were considered equivalent if they were within 200 miles of each other. We used a larger margin for country mentions since our annotators used the GeoNames website online to annotate these locations while our system used the Socrata dataset and there were slight discrepancies between the two sources.

For our secondary study, we annotated only the final location for each record and manually verified whether the final location extracted by our program matched our annotated location for each record. If any part of the correct location was missing or additional FPs were present within

the final location for a record, then we counted it as a single error. This allowed us to estimate the accuracy of our system for other viruses.

2.5 Results

2.5.1 Corpus Statistics

According to the results from our gold standard annotation, 75% of the 5728 GenBank records selected for our influenza case study were found to be insufficient using data from all four fields in the GenBank records. The specificity of the LOIH of 38% of insufficient records was increased using information from the PMC articles. For 90% of these records, it was necessary to read the full-text content of the articles, rather than the abstract only, to make the LOIH more specific.

The percentage of insufficient records in our secondary study was 61%. The specificity of 70% of the insufficient records was increased using information from the full-text content of the PMC articles, primarily tables.

2.5.1.1 Inter-rater Agreement

The IRR for the sufficiency annotation of 2,017 GenBank records from the Influenza case study was found to be 0.984 on average, using Cohen's Kappa statistics as a measure of agreement. Table 1 presents the IRR between each pair of annotators for this task.

	A;B	A;C	B;C
Sufficient;Sufficient	1329	1312	1316
Insufficient;Insufficient	683	684	683
Sufficient;Insufficient	0	17	18
Insufficient;Sufficient	5	4	0
Kappa Value	0.994	0.977	0.980

Table 1: Inter-rater agreement for sufficiency annotation based on Kappa Statistic. A, B and C represent the three annotators respectively

For the final location annotation of 1477 GenBank records from the influenza case study, the IRR was found to be 0.755 on average, using f-score for exact match per record-location linkage

as a measure of agreement between each pair of annotators (the individual f-scores were 0.677, 0.699, and 0.888, respectively).

2.5.1.2 Performance Statistics of the IE System

For determination of the sufficiency of records, our system had a Cohen's kappa value of 0.988 when compared to the gold standard annotation.

Of the 5728 GenBank records used for the influenza case study, our system was able to correctly link 5011 records to the correct final location. For two of the records, neither our system nor our annotators were able to find any geospatial information about their LOIH. For the remaining 715 records, the final location extracted by our system did not exactly match the final location identified by our annotators. However, for 604 of these records, our system output for final location included true positive matches in addition to FPs. For instance, for record GQ463225, our system output for final location was "Guangdong China; Fujian China" while the annotated final location was "Guangdong, China."

Task	Recall	Precision	F-score
Extraction of the location of infected host from GenBank record metadata	0.996	0.953	0.974
Linkage of consistent, text-derived locations of infected host (locations extracted from textual content of related article and consistent with record metadata) to GenBank records	0.800	0.847	0.823
Linkage of consistent, table-derived locations of infected host (locations extracted from tabular content of related article and consistent with record metadata) to GenBank records	0.838	1.0	0.912
Linkage of final locations of infected host (locations produced after integrating geospatial data from record, text and tables) to GenBank records	0.980	0.876	0.925
Mapping of correctly identified final locations of infected host to their correct latitude and longitude coordinates (disambiguation)	0.984	0.948	0.965
Linkage and disambiguation of final locations	0.967	0.832	0.894

Table 2: Recall, Precision and F-score for individual tasks performed by the system

The precision, recall, and f-score of individual tasks performed by the system based on the evaluation criteria described in The Materials and Methods section for the influenza case study is given in Table 2.

The accuracy of our system for linking the 120 records used in our secondary study to their correct final location was found to be 75% (90 records had correct final location).

2.6 Discussion

The results indicate that our system is capable of linking GenBank records to the correct location of sequence collection with a high level of recall and precision. However, since our system evaluation was primarily performed on GenBank records related to the influenza A virus, the results may not be a true reflection of its performance level for GenBank records related to other pathogens. Using the remaining 120 records allowed us to obtain a rough estimate of its accuracy for other viruses but the generalizability of this secondary study is limited by its small sample size. Moreover, at its current state our system is only capable of analyzing records related to PMC articles and therefore, we were limited in our selection of the GenBank records for this study.

Our IRR for sufficiency annotation was very high and matched by the performance of our automated module for this task. However, for final location, we had a relatively low IRR due to misinterpretation of annotation guidelines and missed locations in linked articles, illustrating the difficulty of this task. For instance, one of our annotators listed “New York City, New York, USA” as the final location while the second annotator simply listed “New York, USA” due to missed information in the article. Through repeated annotations, we minimized these mistakes in our gold standard.

For the majority of records whose specificity was increased using information from a related article, it was necessary to retrieve data from the full-text content of the article, including tables. This supports our decision to parse full-text PMC articles rather than PubMed abstracts for this study.

Although we presented evaluation results for the various tasks performed by our system for the influenza case study, we will only present a detailed discussion of its performance in the extraction and disambiguation of final locations, which is the principal objective of the system. Upon

conducting a thorough error analysis for the task of final location extraction, we found that 613 of the 715 records with incorrect final locations were caused by the system's failure to correctly identify the parent ADM1 code of the locations. A total of 605 of these errors were a direct result of the ambiguity of the location "Philadelphia, USA" in GeoNames. Philadelphia was one of the several locations mentioned in a paper linked to over 600 records and our system produced locations such as "Philadelphia, Virginia, USA" and "Philadelphia, New York, USA," respectively, for records containing "Virginia" or "New York" in the GenBank metadata fields. Since we are evaluating the system on a per record-location linkage basis, a single paper associated with a large number of records can have a substantial effect on the system performance. However, by collecting a stratified random sample of papers from the list of all PMC articles related to GenBank records for the influenza A virus, based on the number of records linked to them, we attempted to prevent this from skewing the results significantly (see Appendix D for table showing the distribution of GenBank records and final location errors across the articles)

For 72 records (linked to 8 papers), our system failed to correctly identify the final location due to errors produced by the text linker. Forty-one of these errors (representing 4 papers) were due to the text linker missing relevant relations in the related paper since they did not fit any of the utilized patterns while the remaining resulted from its lack of precision. Although the system's failure to correctly identify the parent ADM1 code of the locations accounted for a significantly greater number of errors in our current evaluation set, the limitations of the text linker, by leading to errors in a larger number of articles, has the potential to be a greater problem in the future depending on the number of records linked to the affected articles.

Spelling errors in the GenBank metadata (e.g., "Jilangsu" recorded in *JN804364* instead of "Jiangsu") and missing location names in GeoNames (e.g., "Pasteur Institute, France") accounted for incorrect final locations in 17 records, none of which had more specific information in the paper. The remaining errors were primarily caused by locations missed by the table linker due to the presence of split cells in the relevant table of a single paper and the failure of the disambiguation module to select the correct candidate for a single ambiguous toponym mention (e.g., "Cambridge").

The coordinates chosen by our disambiguation module were incorrect for a total of seven correctly identified final locations (corresponding to 90 GenBank records) because of the rule-based nature of our program. For five of these locations, the population recorded in GeoNames was *0* and hence our population heuristic had no effect.

The errors in our secondary study primarily resulted from two papers. One of the papers contained more specific information within a table but our program was unable to parse this table since an HTML version of the paper was not available. The second paper utilized two-letter codes to describe the city and state in Brazil from which the virus sample was isolated and our annotator was able to use this data to infer the LOIH of the related sequences. For example, based on the isolate field "CA_SP_P1/0" for record *EU170195,* our annotator was able to deduce that the LOIH for the sequence was Cássia dos Coqueiros, Sao Paolo, Brazil. Our program was unable to make such inferences. The majority of remaining errors were a result of missing locations in GeoNames.

2.7 Conclusion

Our system is capable of linking genetic sequences in GenBank records to the coordinates of their LOIH, using data from the record itself or related PMC articles, with reasonably high accuracy. However, as our error analysis showed, even a single error type can lead to a significant reduction in system performance if a large number of records happen to be affected by this error type. Therefore, as future work, we will attempt to address the different limitations of our system by incorporating additional databases for geographic data such as the Wikipedia dictionary, adding a spell check component to the record location extractor module, and modifying the table linker so that it is capable of parsing more complex tables. In addition, since the rule-based nature of the text linker was a major cause of errors produced by the system, we will test the use of machine learning approaches in this module. Finally, to determine the extensibility of the system, we will evaluate it on other corpora including different species of viruses and other pathogens.

2.8 Contributors

This study was performed under the supervision of G.G., M.S., and G.W. and they all provided significant contributions toward its design and implementation. In addition, G.G. and M.S. helped

considerably with drafting and reviewing the content of the manuscript. T.T. helped with system design and development, evaluated the system based on gold standard data and drafted and revised the manuscript. D.W. helped with system design and development and made significant edits to the manuscript. R.R. performed the stratified random sampling of PMC articles, calculated IRR and made significant edits to the manuscript. R.R., R.B., and M.F. devised the annotation guidelines and annotated the GenBank records. All authors helped review the manuscript.

2.9 Funding

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number R56AI102559 to GG and MS. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

2.10 Competing Interests

The authors have no competing interests to declare.

2.11 References

[1] M. Whirl-Carrillo *et al.*, "Pharmacogenomics knowledge for personalized medicine.," *Clinical pharmacology and therapeutics*, vol. 92, no. 4, pp. 414–7, Oct. 2012.

[2] E. C. Holmes, "The phylogeography of human viruses," *Molecular Ecology*, vol. 13. pp. 745–756, 2004.

[3] D. Magee, R. Beard, M. A. Suchard, P. Lemey, and M. Scotch, "Combining phylogeography and spatial epidemiology to uncover predictors of H5N1 influenza A virus diffusion.," *Archives of virology*, vol. 160, no. 1, pp. 215–24, Jan. 2015.

[4] R. R. Gray and M. Salemi, "Integrative molecular phylogeography in the context of infectious diseases on the human-animal interface.," *Parasitology*, vol. 139, no. 14, pp. 1939–51, Dec. 2012.

[5] R. Hovmöller, B. Alexandrov, J. Hardman, and D. Janies, "Tracking the geographical spread of avian influenza (H5N1) with multiple phylogenetic trees," *Cladistics*, vol. 26, no. 1, pp. 1–13, Feb. 2010.

[6] D. A. Janies *et al.*, "Phylogenetic visualization of the spread of H7 influenza A viruses," *Cladistics*, vol. 31, no. 6, pp. 679–691, 2015.

[7] J. Chan, A. Holmes, and R. Rabadan, "Network analysis of global influenza spread.," *PLoS computational biology*, vol. 6, no. 11, p. e1001005, Jan. 2010.

[8] P. Elliott and D. Wartenberg, “Spatial epidemiology: current approaches and future challenges.,” *Environmental health perspectives*, vol. 112, no. 9, pp. 998–1006, Jun. 2004.
[9] D. A. Benson *et al.*, “GenBank,” *Nucleic Acids Research*, vol. 41, 2013.

[10] J. Raghwani *et al.*, “Endemic dengue associated with the co-circulation of multiple viral lineages and localized density-dependent transmission.,” *PLoS pathogens*, vol. 7, no. 6, p. e1002064, Jun. 2011.

[11] N. R. Faria, M. A. Suchard, A. Rambaut, and P. Lemey, “Toward a quantitative understanding of viral phylogeography.,” *Current opinion in virology*, vol. 1, no. 5, pp. 423–9, Nov. 2011.

[12] D. Janies, A. W. Hill, R. Guralnick, F. Habib, E. Waltari, and W. C. Wheeler, “Genomic analysis and geographic visualization of the spread of avian influenza (H5N1).,” *Systematic biology*, vol. 56, no. 2, pp. 321–9, Apr. 2007.

[13] “GeoNames.” [Online]. Available: http://www.geonames.org/. [Accessed: 13-May-2015].

[14] R. Bossy, W. Golik, Z. Ratkovic, P. Bessières, and C. Nédellec, “BioNLP shared Task 2013 – An Overview of the Bacteria Biotope Task,” in *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, 2013, pp. 161–169.

[15] J. Tamames and V. de Lorenzo, “EnvMine: a text-mining system for the automatic extraction of contextual information.,” *BMC bioinformatics*, vol. 11, p. 294, Jan. 2010.

[16] I. N. Sarkar, “Leveraging Biomedical Ontologies and Annotation Services to Organize Microbiome Data from Mammalian Hosts,” *AMIA Annu Symp Proc*, vol. 2010, pp. 717–721, 2010.

[17] E. S. Chen and I. N. Sarkar, “Towards Structuring Unstructured GenBank Metadata for Enhancing Comparative Biological Studies,” *AMIA Jt Summits Transl Sci Proc*, vol. 2011, pp. 6–10, Mar. 2011.

[18] E. S. Chen and I. N. Sarkar, “MeSHing molecular sequences and clinical trials: a feasibility study.,” *Journal of biomedical informatics*, vol. 43, no. 3, pp. 442–50, Jun. 2010.

[19] H. Miller, C. N. Norton, and I. N. Sarkar, “GenBank and PubMed: How connected are they?,” *BMC research notes*, vol. 2, no. 1, p. 101, Jan. 2009.

[20] O. Selama, P. James, F. Nateche, E. M. H. Wellington, and H. Hacène, “The world bacterial biogeography and biodiversity through databases: a case study of NCBI Nucleotide Database and GBIF Database.,” *BioMed research international*, vol. 2013, p. 240175, Jan. 2013.

[21] T. Tahsin *et al.*, “Natural Language Processing Methods for Enhancing Geographic Metadata for Phylogeography of Zoonotic Viruses,” *AMIA Jt Summits Transl Sci Proc*, vol. 2014, pp. 102–111, Apr. 2014.

[22] “Xpdf: Download.” [Online]. Available: http://www.xpdfreader.com/download.html. [Accessed: 04-Mar-2014].

[23] E. Sayers, *E-utilities Quick Start*. National Center for Biotechnology Information (US), 2013.

[24] M. D. Lieberman, H. Samet, and J. Sankaranarayanan, “Geotagging with local lexicons to build indexes for textually-specified spatial data,” in *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, 2010, pp. 201–212.

[25] S. Ladra, M. R. Luaces, O. Pedreira, and D. Seco, “A Toponym Resolution Service Following the OGC WPS Standard,” in *Web and Wireless Geographical Information Systems*, 2008, vol. 5373, pp. 75–85.

[26] “Country List ISO 3166 Codes Latitude Longitude | Socrata.” [Online]. Available: https://opendata.socrata.com/dataset/Country-List-ISO-3166-Codes-Latitude-Longitude/mnkm-8ram. [Accessed: 13-May-2015].

[27] M. Scotch *et al.*, “Enhancing phylogeography by improving geographical information from GenBank.,” *Journal of biomedical informatics*, vol. 44 Suppl 1, pp. S44–7, Dec. 2011.

[28] A. X. Chang and C. D. Manning, “SUTIME: A Library for Recognizing and Normalizing Time Expressions,” in *LREC*, 2012.

[29] D. Weissenbacher *et al.*, “Knowledge-driven geospatial location resolution for phylogeographic models of virus migration.,” *Bioinformatics (Oxford, England)*, vol. 31, no. 12, pp. i348–i356, Jun. 2015.

[30] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, “GATE: A framework and graphical development environment for robust NLP tools and applications,” 2002.

[31] Y. Tsuruoka and J. Tsujii, “Bidirectional inference with the easiest-first strategy for tagging sequence data,” in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT ’05*, 2005, pp. 467–474.

[32] J. L. Leidner, “Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names,” University of Edinburgh, 2007.

[33] G. Hripcsak and A. S. Rothschild, “Agreement, the f-measure, and reliability in information retrieval.,” *Journal of the American Medical Informatics Association : JAMIA*, vol. 12, no. 3, pp. 296–8, Jan. 2005.

s

2.12 Appendix A: Detailed description of corpus selection

The first step in this study involved downloading 309,807 GenBank records related to the Influenza A virus (NCBI taxonomy id of 197911) using the following query:

http://www.ncbi.nlm.nih.gov/nuccore/?term=txid197911%5BOrganism:exp.

Since a major objective of our project is to improve the specificity of the location metadata of GenBank records by extracting information from related full-text PubMed articles, we filtered these records based on whether they had a PMID linked to them. This produced a set of 102,949 records which were linked to 1424 unique PMIDs. For the scope of this project, we concentrated specifically on papers that were part of the PMC database and used the PMID to PMCID convertor online (to narrow down the list of related articles to 598 papers with valid PMCIDs, corresponding to 62,103 GenBank records. 10% of this set of papers, representing 60 PMC full-text articles, were selected via stratified random sampling to create our evaluation corpus. The stratified random sampling was performed by first dividing the 598 papers into 6 strata based on the number of GenBank records related to them and using the Excel rand() function for randomly selecting 10% of the papers from each strata for inclusion in our study; this was done since the number of GenBank records related to a given article may vary widely, ranging from just 1 to over a 1000. The first stratum contained papers with 1-4 records, the second 5-8, the third 9-24, the fourth 25-76, the fifth 76-500 and the sixth 501+. The final set of 60 papers were linked to 5,728 GenBank records. We manually downloaded the PDF versions of each of these 60 papers from the PubMed website and wrote a script utilizing the freely available pdf-to-text tool to convert each of them into text files. For downloading relevant fields from the selected GenBank records, we used the NCBI Entrez Utilities API. The following fields were downloaded for each record: ‘Country’, ‘Strain’, ‘Organism’, ‘Isolate’, ‘Date’, ‘Gene’, and ‘Host’.

2.13 Appendix B: Patterns used for text linkage

For every toponym detected by our toponym detector, our text linkage extraction algorithm first checks whether the sentence containing the toponym matches the very general “Location Pattern”

rule in Table 3. If it matches the pattern, then our algorithm proceeds to analyze the sentence to determine whether it can be extracted using any of the remaining patterns listed in Table 3.

“Isolation” Pattern	“.* (isolated|collected) .* (in|from) .*”+location+”.*”
	“.* we .* (collect|isolate) .* (in|from) .*”+location+”.*”
	“.* (isolation|collection) .* (in|from) .*”+location+”.*”
“Location” Pattern	“.* (in|from|at) .* ”+location+”.*”
“Our study” Pattern	“.*we used.*”
	“.*current study.*”
	“.*in this study.*”
	“.*our study.*”
	“.*we examined.*”
	".*we studied.*"
“Metadata” Pattern	“.* ”+host+”.*”
	“.* ”+date_of_collection+”.*” (either year, month or entire date)
	“.* ”+strain+”.*”
	“.* ”+virus_name+”.*”
	“..* ”+accession+”.*”

Table 3: Rules used for linkage of locations detected in textual content of related article to Genbank records

2.14 Appendix C: annotation

The annotation of the 5.728 GenBank records was performed in an excel sheet with columns labeled “Sufficient?”, “Record Locations”, “Final Record Locations”, “Text Locations”, “Table locations”, “Final Location”, “Latitude” and “Longitude”. The approach used for annotating each of these columns, along with the purpose of each annotation, are given below:

1. Sufficient: This indicates whether or not a record is sufficient based on geospatial data from all relevant fields of the GenBank records and was used to evaluate the sufficiency determination task. We looked at the ‘Country’ field, the ‘Strain’ field, and the ‘Organism’ field, in the order stated, for more specific geographic information, until the combined

information made the record sufficient or the information from all four fields failed to make it sufficient. It can have a value of either 'Yes' or 'No'. To determine sufficiency, we retrieved the feature codes of geospatial locations using the GeoNames website. If a place could have several different feature codes, we erred on the side of caution and chose the less specific feature code. For instance, the metadata of GenBank record CY053893 contained 'Buenos Aires, Argentina' and although Buenos Aires can be either a province or city in Argentina, we assumed that it referred to the province and labeled the record as insufficient.

2. Record Locations: This includes a list of all locations mentioned within the 'Country', 'Strain', and 'Organism' fields of the record, separated by semi-colons. It was used to evaluate the record location extraction task.
3. Text Locations: This lists all unique mentions of geospatial locations related to a record which the annotators found in the paragraphs of linked papers. It was used to evaluate the set of locations extracted by the Text Linker and found to be consistent with record location by the Data Integrator. This column was only annotated for insufficient records. The listed locations were often redundant (e.g. New York and USA will be listed as two separate locations although New York is a part of USA) and not useful (they may have already been present in the metadata of the records). However, they represent the unique location mentions which we expect our system to be able to find in the textual content of the paper.
4. Table locations: This lists all unique mentions of geospatial locations related to a record which the annotators found in the tables contained within linked papers. It was used to evaluate the set of locations extracted by the Table Linker and found to be consistent by the Data Integrator. As in case of the text annotation, this column was only annotated for insufficient records.
5. Final location: This states the final location (s) for a record based on combined information from all fields in the GenBank record and the textual and tabular content of the related paper, and was used to evaluate the final location output from the data integrator module of our system. This column may contain more than one distinct location in cases where a paper was associated with multiple records and it was not possible to identify which of the

locations of isolation mentioned in it belonged to the specific record. Each distinct location can include either a single country name (e.g. 'China'); an ADM1 location and its parent country, separated by commas (e.g. 'California, USA'); a location more specific than ADM1 along with its parent ADM1 location and parent country location, separated by commas (e.g. Sydney, New South Wales, Australia); or a location more specific than ADM1 along with its parent country location (e.g. 'Shantou, China'), separated by commas. Individual locations are separated by semi colons.

6. Latitude and Longitude: This states the latitude and longitude coordinates of the final location of isolation extracted for a record. It was used to evaluate both the location disambiguation task as well as the ability of our system to link each record to the correct set of latitude and longitude coordinates. We searched the GeoNames website for these coordinates and chose the one with the least specific feature code. In cases where GeoNames did not a find a match for our query, we used Google Maps.

2.15 Appendix D: Distribution of records and program errors for final location across PubMed articles used in the influenza case study

PubMed ID	Number of related records	Number of related records for which the extracted final locations contained error(s)
20943966	627	610
24244615	64	25
22470427	24	24
23458714	30	8
23441208	52	7
2041090	39	7
18704172	7	7
20167132	5	5
21645421	5	5
21930918	12	4
23285143	761	2
16333111	146	2
21490925	4	2
6296449	2	2

18214200	2	2
22718569	28	1
2349225	1	1
6253883	1	1
17652402	1648	0
21047959	440	0
23029335	296	0
20975994	267	0
20610681	240	0
20592108	176	0
22050764	114	0
18394282	96	0
16824249	88	0
20202225	70	0
20875285	64	0
22012020	53	0
11371620	37	0
21637809	36	0
16306617	28	0
20631138	24	0
21900171	24	0
23920350	24	0
23650608	20	0
22984519	17	0
18032512	16	0
19359528	16	0
17881439	13	0
24086762	11	0
21392430	9	0
19380727	8	0
22279582	8	0
22709385	8	0
22733885	8	0
22879613	8	0
23087121	8	0
23580714	8	0
23950136	8	0
21173241	7	0
1731092	6	0

12857911	4	0
11158130	3	0
20202440	3	0
2780295	1	0
6828387	1	0
20220153	1	0
23042757	1	0

Table 4: Distribution of records and program errors for final location across PubMed articles used in the influenza case study.

3 NAMED ENTITY LINKING OF GEOSPATIAL AND HOST METADATA IN GENBANK FOR ADVANCING BIOMEDICAL RESEARCH

Authors: Tasnia Tahsin*, Davy Weissenbacher*, Demetrius Jones-Shargani, Daniel Magee, Matteo Vaiente, Graciela Gonzalez, Matthew Scotch

**The two authors contributed equally and are co-first authors. Ordering is based on the alphabetical ordering of their last names*

3.1 Abstract

GenBank is a popular NCBI database for submission and analysis of DNA sequences for biomedical research. The resource is part of the Entrez environment which enables for cross-linking of concepts and entries in other participating NCBI databases such as Taxonomy, PubMed, and Protein. For example, a GenBank record of an influenza A hemagglutinin gene DNA sequence might have a link to the Taxonomy database for the organism, a link to the related article in PubMed (if published), and a link to the Protein entry for the hemagglutinin protein. Despite its importance in biomedical research such as population genetics, phylogeography, and public health surveillance, the host and geospatial metadata of genetic sequences in GenBank are not linked to any database. Therefore, to facilitate biomedical research based on georeferenced DNA sequences and/or DNA sequences with normalized host names, we designed and developed a framework that enriches GenBank entries by linking their host metadata to the NCBI Taxonomy database and their geospatial metadata to a comprehensive knowledge base of geographic locations called GeoNames. Here, we introduce a database created through the application of this framework to virus sequences in GenBank and evaluate our normalization algorithms on a set of manually annotated records pertaining to viruses. Although currently applied to viruses, our framework can be easily extended to other organisms, and we discuss the potential utilization of our resource for biomedical research. The developed database is available for download at https://tinyurl.com/GeoHostDB. An online interactive version of the database is also available at https://zodo.asu.edu/zoophydb/. The github repository for the source code of our framework is available at https://tinyurl.com/GenbankFactory.

3.2 Introduction

GenBank is a public database of nucleotide sequences developed and maintained by the National Center for Biotechnology Information (NCBI), which is part of the U.S. National Library of Medicine (NLM) of the National Institutes of Health (NIH) [1]. With its participation in the International Nucleotide Sequence Database Collaboration (INSDC), NCBI exchanges sequences with international institutes such as the European Nucleotide Archive (ENA) [2] and the DNA Data Bank of Japan (DDBJ) [3]. At the time of writing, GenBank contains a total of 200,877,884 sequences [4], along with pre-defined metadata describing each sequence. Over two million of these sequences are of virus origin, and include metadata such as the name of infected host of the virus, the location of infected host of the virus (LOIH), and the name of the gene the sequence corresponds to.

Viruses represent one of the principal causes of emerging and re-emerging infectious diseases across the world [5], and, therefore, understanding their evolutionary dynamics and geographical transmission, through diverse methods of analysis, is of critical importance. As one of the most comprehensive sources of virus sequence information, GenBank presents an invaluable resource for a wide range of virus-related research. It is frequently used in fields such as phylogenetics, phylogeography, molecular epidemiology, evolutionary biology, and environmental health for studying viruses through a variety of different approaches. In addition to genetic sequence data, the rich metadata present in many GenBank records are vital for analysis and comparison. For instance, when mapping the global spread of each type of Dengue viruses across a time span of 70 years, Messina *et al*. extracted the type and geographical coordinates of 1,070 GenBank records pertaining to Dengue viruses from their respective metadata fields [6]. Similarly, Scotch *et al*. also utilized the geospatial metadata available for GenBank records when conducting a phylogeographic analysis of Influenza A H5N1 viruses isolated from Egypt [7].

One significant challenge faced by researchers in their efforts to incorporate GenBank metadata within their study, is the task of appropriately normalizing the data so that it is usable. Although GenBank contains distinct fields for storing sequence-related metadata, it does not place strict constraints on values that an author may enter for each field. As a result, many of the metadata

fields in GenBank are semi-structured in nature and must be processed before being utilized by a researcher. For instance, the host field of GenBank records with accession numbers AB618040 [8], AB618529 [9] and AJ312308 [10] contains the values “Homo sapiens”, “Homo sapiens 54 years old female” and “Man”, respectively, to denote the same species. Therefore, if a researcher intends to focus on virus sequences infecting, for example, humans and chimpanzees, they would first have to guess the different possible ways of denoting human and chimpanzee hosts, then query the GenBank website for each such possibility, and finally normalize each host field manually to allow grouping based on its value.

When extracting geographic metadata denoting the location of the infected host (LOIH) of a virus sequence, researchers may frequently have to perform an additional step of integrating geographic information from different fields in the GenBank record, prior to normalization. The designated field for storing the LOIH of sequences in GenBank is called the *country* field. Despite its name, the country field may contain geographic metadata of varying degrees of specificity, rather than only country-level information. For instance, the annotated data in the country field of the GenBank record with accession number CY045959 is "Canada: Ontario" [11]. Due to the specific nature of virus nomenclature, additional geographic information may often be found in the strain field and isolate field of GenBank records. For instance, the annotated data in the strain field of this record is "A/Toronto/T5294/2009(H1N1)" [11]. Combined with the information in the country field, it can be inferred that the LOIH of the virus is "Toronto, Ontario, Canada". This process of extracting, integrating and normalizing the LOIH of sequences from GenBank record metadata can be highly challenging, especially when a researcher is not very familiar with the geographic region in which the study is being conducted. The ambiguous nature of many locations can make this process even more difficult. For instance, the location “Malang, Indonesia” may be mapped to 20 distinct geo-coordinates based on GeoNames [12], a comprehensive database of geographic locations across the world. In 2005, GenBank introduced the lat_lon field [13] which, in the case of viruses, may be used to store the specific latitude and longitude coordinates of their LOIH. However, in our review, we found that this field is missing in over 99% of all GenBank records pertaining to viruses.

Therefore, for the large majority of GenBank records, the task of geocoding is left to each individual researcher.

In this study, we describe the design and development of an integrated framework for normalizing host and location metadata in GenBank records pertaining to viruses. We applied a rule-based framework to map the name and location of the infected hosts of viruses to their corresponding NCBI taxonomy IDs [14] and GeoNames IDs [12] respectively. Our algorithm successfully linked 1,971,328 GenBank records to the GeoNames database, and 1,592,541 GenBank records to the Taxonomy database based on their host names. Prior to normalizing the LOIH of virus sequences in GenBank, we first used an automated approach to integrate data from different fields in the record which may contain geographic metadata. Therefore, our database includes the most comprehensive geographic metadata denoting the LOIH of each virus sequence in GenBank, which our algorithm is capable of extracting. To the best of our knowledge, this is the first framework that normalizes these two types of GenBank metadata for all virus-related GenBank records.

Given the significance of normalized GenBank metadata in a wide range of virus-related studies, our framework would help support a variety of different approaches used for understanding and/or analyzing virus epidemiology, migration patterns, and evolutionary dynamics. This, in turn, may lead to major advances in infectious disease surveillance, and vaccine design and distribution, thereby enhancing our ability to control and contain disease outbreaks. In addition, our normalization algorithms linked each GenBank metadata to widely used and well-managed databases. This would facilitate cross-database queries, allowing the conduction of many new analytical studies. Moreover, the methods of normalization described here may also be easily applied to create similar databases for organisms such as bacteria or eukaryotes. Therefore, the work presented in this paper has the potential to considerably accelerate research in diverse biomedical fields.

3.3 Related Work

Over the past few years, NCBI has undertaken several large-scale efforts to add more structure to its data, resulting in the development of valuable resources such as BioSample [15], [16], Refseq

[17], [18], NCBI Virus Variation [19], [20] and NCBI Viral Genomes [21], [22]. These resources facilitate curation of GenBank metadata and are crucial for advancing biomedical research. However, we believe that our framework is distinctly different from each of them and serve a purpose not yet satisfied by any existing resource that we are aware of. The BioSample project represents a significant attempt by NCBI to integrate data across different resources, and provides an intuitive interface to facilitate submission of rich and consistent metadata. However, it relies on manual submission of metadata and is not linked to a large section of virus GenBank records. The Refseq database is a widely used resource within the research community which includes non-redundant, well-annotated genetic sequences but it requires manual curation of data, and, once again, a large portion of virus GenBank records do not have Refseq links. The NCBI Virus Variation project, which is part of the NCBI Viral Genomes project, utilizes a semi-automated pipeline for mapping GenBank metadata, including host and geographic metadata, to a controlled vocabulary. However, the pipeline is currently applied to newly-released GenBank records pertaining to seven viruses only. In contrast, we have successfully applied our automated system to over two million GenBank records pertaining to viruses. Moreover, the pipeline used by the NCBI Virus Variation project appears to map the geographic metadata of virus records to their corresponding countries/continents/regions only to allow recognition of up-to country-level hierarchy, while our framework normalizes the metadata to specific GeoNames entries (which includes their geographic coordinates) and is capable of recognizing up-to state/province-level hierarchy. Also, unlike our system, the NCBI Virus Variation pipeline appears to map host names to a controlled vocabulary of taxonomic host groups rather than specific taxonomy ids.

Although this work represents the first effort to create a comprehensive database including the normalized forms of the infected host and LOIH of all virus sequences in GenBank, several attempts have been previously made to normalize different GenBank metadata fields for different organisms. In our prior work [23], we used a rule-based approach similar to the one described here to extract, integrate, and normalize the LOIH of virus sequences in GenBank. However, instead of applying our approach to develop a database of virus-related GenBank records with normalized LOIH, we used it to develop a system for enhancing existing geographic metadata in “insufficient” virus-

related GenBank records by extracting additional information from linked full-text publications. We defined “Insufficiency” as geographic metadata which was not more specific than Administrative Division 1 (ADM1) level i.e. state or province level. For instance, “Arizona, USA” would be categorized "insufficient" while “Maricopa County, Arizona, USA” would be categorized "sufficient". Therefore, once our system found “sufficient” geographic metadata in a GenBank record, it would stop searching. For instance, if the geographic metadata in a record was “Tempe, Maricopa County, Arizona, USA”, our system would stop searching once it found “Maricopa County”, thereby missing the more specific location “Tempe”. Here, we updated our algorithm so that it finds the most specific geographic location, along with its parent ADM1 and country-level location, if present, for semantic context. Therefore, in the previous example, our current system would extract, and subsequently normalize, “Tempe, Arizona, USA”. Moreover, the rules for LOIH extraction in the system developed through our prior work were primarily designed for GenBank records pertaining to only the influenza virus. For this study, we added rules to optimize geographic metadata extraction for non-influenza viruses as well, and introduced additional features, such as a simple Lucene-based spell corrector, to minimize errors for all organisms. Furthermore, in our prior work, we used an SQL database for storing and querying the GeoNames knowledge base. Here, we migrated to a Lucene index representation of the knowledge base to enable faster queries.

In another recent work, Gratton *et al*. [13] utilized an automated approach for geocoding all previously un-geocoded GenBank records associated with tetrapods. However, they did not extend their study to include viruses, and limited the extraction of geographic metadata from the *country* field only, while we integrated geographic metadata from different fields in virus-related GenBank records for this study. Furthermore, they mapped the extracted geographic metadata to their respective latitude and longitude coordinates, while we mapped sequences to their corresponding GeoNames IDs, whenever possible, in addition to their geo-coordinates. This would enable cross-database studies involving the GeoNames database and the GenBank database, and provide a unique, normalized string representation of each LOIH to facilitate studies such as discrete phylogeography, where each LOIH is represented as a discrete character state [24].

Recent efforts have also been made to extract and normalize non-geographic metadata in GenBank. For instance, Sarkar [25] extracted the anatomical source of microbiome bacteria in ten mammalian hosts from the *isolation_source* and *note* fields in GenBank records, and normalized them using existing ontologies and annotation services available through the National Center for Biomedical Ontologies (NCBO) [26]. In a separate work, Chen and Sarkar [27] conducted a feasibility study for normalizing the *host* and *isolation_source* fields in GenBank. They applied an automated approach to normalize *host* fields to their corresponding Taxonomy IDs, and used the NCBO web service annotator for normalizing the *isolation_source* field based on different ontologies. However, their work only involved an exploratory analysis of GenBank records, and no datasets including the normalized fields was made publicly available. Another related work by Sinclair *et al*. [28] introduced Seqenv, a software for linking genetic sequences to the Environmental Ontology [29]. However, Seqenv takes genetic sequences as input instead of GenBank records, and the linking is performed based on the *isolation_source* field in GenBank. Therefore, it is specifically geared towards assisting researchers specializing in environmental genomics while our framework serves as a general framework for normalizing GenBank metadata, which may address the needs of diverse research areas.

Research in the emerging domain of *viroinformatics* has also lead to the development of many computational tools and databases to support the work of virologists. Sharma *et al.* [30] provided an exhaustive list of such resources, and included key features and functions of each resource. However, none of the listed resources were reported to have used computational methods for normalizing the host and geospatial metadata of all virus sequences in GenBank.

Outside of GenBank and viral genomics, different normalization methods have been utilized in a wide range of studies to normalize mentions found in free-text articles [31]–[35], tables and lists in web documents [36], [37], social media [38] and other databases/knowledgebases [39]. Although we exploited the basic principles involved in some of these normalization techniques, which are commonly used by researchers, the exact heuristics applied here remain unique to our study.

3.4 Methods

Our study can be divided into three distinct stages: 1) Database Design and Development 2) Entity Normalization 3) Evaluation. Below, we describe each stage in detail.

Database Design and Development

3.4.1.1 Database Design

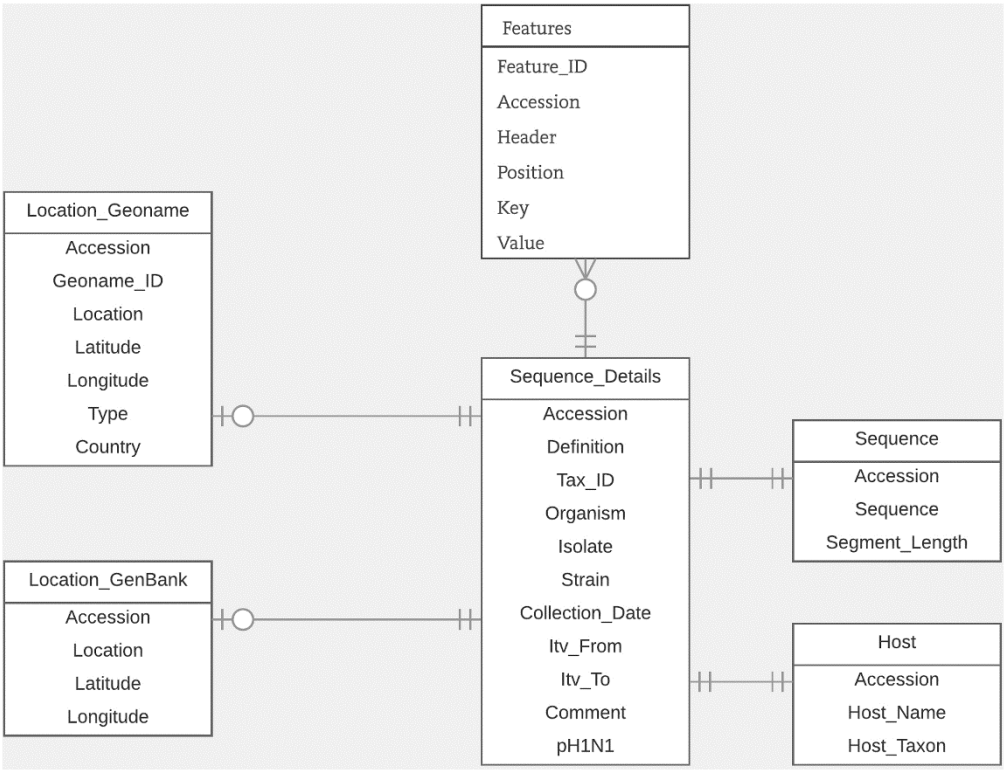


Figure 1: Database Schema

For this study, we designed an efficient and flexible database schema. In Figure 1 we illustrate the portion of the schema relevant to the task of entity linking. Within our database, the GenBank accession number is the main identifier used to connect all related metadata for each virus sequence. We organized the database around the “Sequence_details” table which includes sequence metadata extracted from important fields in GenBank records such as the organism field, isolate field, strain field, collection_date field, etc. We stored the data extracted from the host field, along with their normalized forms, in the “Host” table. We stored the data from the country field,

along with the latitude and longitude coordinates (which we derived from the lat_lon field), in the “Location_GenBank” table. In the table “Location_Geoname”, we saved the integrated LOIH which we extracted from the relevant fields in each GenBank record. In this table, we also stored its corresponding GeoNames ID and latitude and longitude coordinates. We chose to use separate tables for storing the normalized host name and LOIH of each virus sequence to facilitate updating and/or analyzing the novel pieces of information derived through this study. We stored additional metadata from the “Features” section of each GenBank record in the “Features” table. Here, we utilized a flexible structure by using “Key” and “Value” columns to store each feature. Finally, we stored the entire nucleotide sequence included in each GenBank record in the “Sequence” table.

3.4.1.2 GenBank Data Download

NCBI offers several web-based services to access or download the entire GenBank database. Here, we used the anonymous ftp server located at https://www.ncbi.nlm.nih.gov/genbank/ftp/ to acquire all GenBank records pertaining to viruses listed in the gbvrl files (excluding laboratory strains). Using a parser written in Java, we sequentially downloaded all GenBank flat files corresponding to virus nucleotide sequences from the anonymous ftp server. After downloading each file, we ran our parser to automatically extract relevant data for each sequence contained in the file and stored them in our SQL database.

3.4.2 Entity Normalization

The task of normalization aims to map the mention of a concept to its corresponding ID in a predefined knowledge base (KB) [40]. For example, in the sentence "one SOR strain that was also isolated from a human in Germany" [41], p. 2052, the mention "human" can be linked to the concept Homo sapiens (ID:9606) in the NCBI Taxonomy [42] and the mention "Germany" to the concept of Federal Republic of Germany (ID:2921044) in the GeoNames database [43]. The normalization task is also known as "concept mapping", "concept grounding" or, as in our study, "entity linking" when the concepts are only limited to entities. In entity linking, the mention of concepts, such as quality, process, or events, are excluded from normalization.

Normalizing concepts in documents is made difficult due to the presence of various linguistic phenomenon. An intuitive approach to normalize the mention of a concept appearing in a document is to compare the mention with each entry in the chosen KB. If an entry matches exactly with the mention, the ID of the entry is linked to the mention. However, synonyms, polysemy, acronyms, and spelling variations render a search by exact match ineffective [40].

When exploring the feasibility of normalizing concepts in the semi-structured *host* and *isolation_source* fields of the GenBank database, Chen *et al*. [27] noted that the *host* field often included the common names of the host, rather than their scientific names, in a wide range of different formats, along with additional information about the host, such as its age and gender (*e.g.* for accession CY138679 [44], the field *host* is "American black duck; gender M; age L - Local"). The *isolation_source* field presented an even richer syntactic and semantic diversity since its values varied based on the anatomy of the hosts. Therefore, in both cases, complex methods are required to successfully normalize the fields, and a simple search by exact match would likely be ineffective. Complex approaches of entity normalization rely on the exploitation of the properties of mentions and concepts, along with the contexts in which they appear [45]. Below, we list some of the common features used in such complex approaches for entity normalization:

- Names similarities: The most intuitive and commonly used property to link a concept to a mention is the similarity between their names. When a strict string matching is not directly applicable, a distance of some sort may be computed between the string of the mention and the names of different concepts in the KB, to search for the closest concept.
- Concept popularities: Some concepts are more frequently used than others. For example, if the name “Marie Currie” is mentioned in a document, it is more likely to refer to the famous Polish physicist than the less famous American rock singer, “Marie Michelle Currie”. A simple metric to confirm this claim may be derived by comparing the number of Wikipedia articles referring to the physicist with the number of articles referring to the rock singer. An *a-priori* probability can model this likelihood and be used to bias the default choice of a concept for a given mention.

- Lexical Context: When normalizing the mentions in a document, it may often be possible to exploit the lexical context around each mention (*e.g.* the words in the paragraph containing the mention) by comparing it with the lexical context of all possible concepts in the KB (*e.g.* the words describing the concept in the KB). The lexical context of the concept which corresponds to the mention is expected to be more similar to that of the mention in the document. However, when normalizing concepts in the fields of a database, such context may not always exist or be very informative. For instance, the host name entered in the GenBank record with accession KR349276 [46] is "mouse". This mention is ambiguous with the taxonomy concepts Shrew mouse (ID: 10093) [47], House mouse (ID:10090) [48], and Western Wild Mouse (ID:10096) [49]. However, it is not possible to exploit lexical context to disambiguate this mention since it is not surrounded by any other word in this field.
- Semantic Context: The concepts discovered in a document are rarely independent of each other, and the chosen concept for a mention should be coherent with the concepts chosen for other mentions in the document. In our previous example, if the name “Marie Currie” is found in a document mentioning the names “Cherie Currie” and “Steve Lukather”, which match the names of the American rock singer Marie Michelle Currie's sister and husband respectively, it is more likely to refer to the American rock singer rather than the more famous Polish physicist. Therefore, the semantic context of a mention may often be used to successfully disambiguate the entities in a document.

In this study, we used several of the entity linking strategies listed above for normalizing GenBank metadata, in addition to using search by exact match. When normalizing the *host* field, we exploited the name similarities between mentions and concepts, as well as the popularity of the concepts. In case of geospatial metadata normalization, we exploited the semantic context of the mentions along with the name similarity and concept popularity features. Geographic locations are hierarchical in nature and GenBank metadata often includes hierarchical information for the location of infected host of a virus, which we refer to here as semantic context. For instance, if the “country” field of a GenBank record contains “Paris, Texas, USA”, our algorithm would use “Texas” and “USA”

as semantic context when disambiguating “Paris”. As a result, “Paris”, in this specific case, would be mapped to the GeoNames ID of 4717560 [50], representing the city of Paris in Texas, USA, rather than the GeoNames ID of 2988507 [51] representing the capital city of France, which is the more widely known of the two locations. Without taking the semantic context of “Texas,USA” into consideration, our algorithm would have mapped Paris to the capital city of France. We did not utilize lexical context in either of the two normalization algorithms presented here since the description included in GenBank for each metadata is too short to benefit from this normalization strategy. Further details about each normalization method are described below.

3.4.2.1 Host Normalization

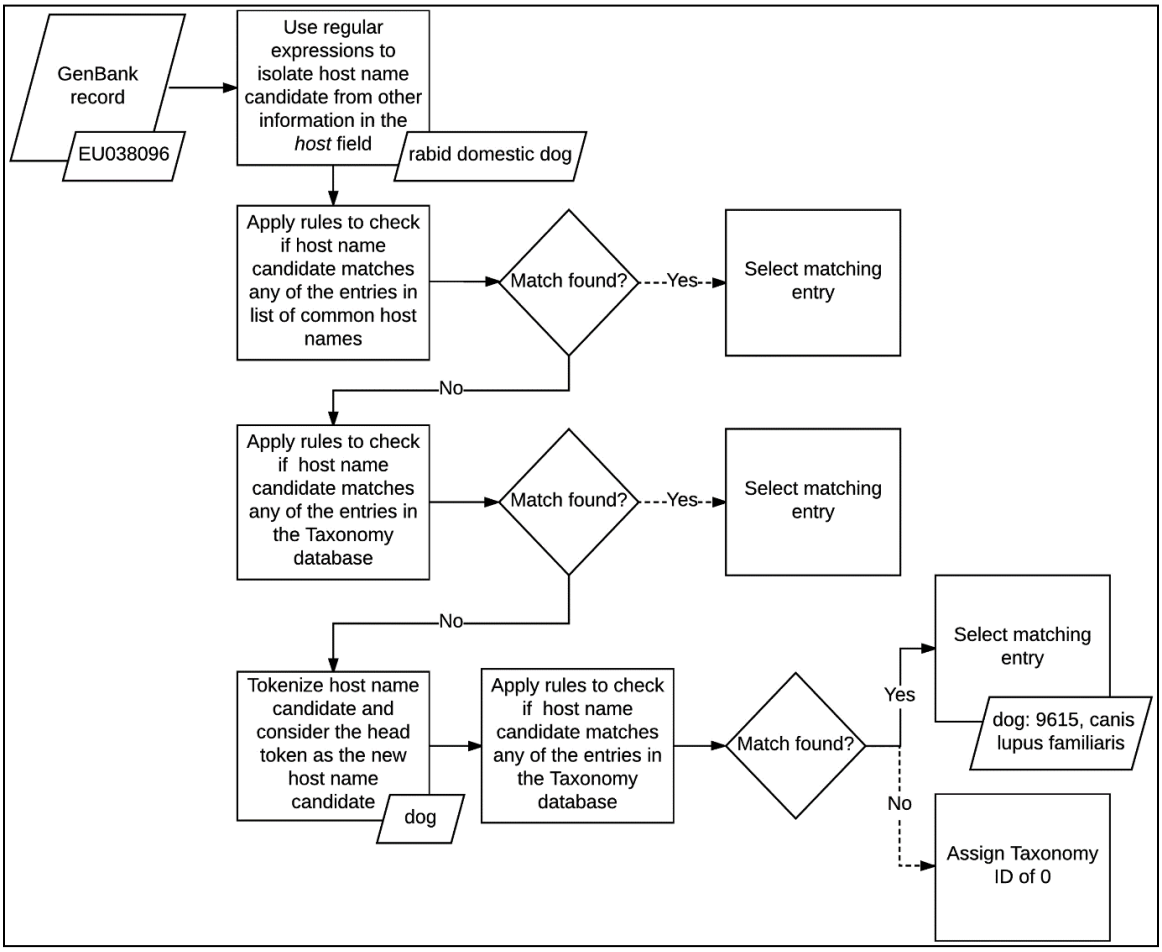


Figure 2: Host metadata extraction and normalization algorithm

We normalized the host field in GenBank records by applying a set of matching rules in sequence (see Figure 2). First, we isolated the name of the host from any additional information the field may contain using a series of handwritten regular expressions. The regular expressions we applied were designed to recognize several formats followed by authors when entering this field during the sequence submission process. For example, based on one of our rules, we discarded any text in the field which followed the occurrence of the first punctuation mark, if the punctuation was not a period, and kept only the remaining phrase. For example, in GenBank record KT390491: "Abelmoschus angulosus; IC-140156" [52], we only kept *Abelmoschus angulosus*.

Once we isolated the names of the hosts, we applied a set of rules to map the mention of common host names, such as mouse and human, to their corresponding IDs in the Taxonomy database. If none of these rules matched, then we implemented a second set of rules to search for regular patterns against the entire taxonomy tree, instead of a small set of common hosts. If no matching host name was found, we tokenized the name and checked if the head token matched any of the rules included in the second set. If a match was still not found, then we assigned the *host* field a Taxonomy ID of 0, to indicate that the host name of the record was unknown.

Although several NLP tools currently exist for the generic task of species normalization [34], [53], we opted to develop our own algorithm for this domain instead of adopting one of the existing tools. Most GenBank records pertaining to viruses contain very short descriptions of the infected host within the host field. In many cases the included host name is a scientific name which can be directly mapped to an NCBI Taxonomy entry. Non-scientific host names used typically fall within a limited set of common host names. Therefore, we attempted to use a simple rule-based approach for normalizing the host names in GenBank records rather than applying more complex NLP tools. This allowed us to keep our methods as simple and efficient as possible while still having complete flexibility to make any changes needed to enhance performance specifically for this domain.

3.4.2.2 Geospatial Metadata Normalization

To extract and normalize the geospatial metadata of virus sequences in GenBank, we constructed a Lucene index of geographic locations, based on the GeoNames database, to serve as our knowledge base of location names. The GeoNames database, which encodes the properties

and hierarchical structure of over 10 million geographic locations, is a widely-used resource for geographic information extraction. However, it contains many entries such as “rat” and “fox” which may generate many false positives. Therefore, we collected different lists of commonly used words from different sources to filter them out. This includes a list of the names of common virus hosts and a list of English stop words [23]. GeoNames also includes the alternate names of each location in different languages. We included these alternate names in our knowledge base for all ADM1-level locations to maintain a high recall. For country-level locations we manually added commonly used country names and considered the Socrata dataset [54], which includes geospatial data for 243 countries, when adding these alternate names [23]. For all other locations, we did not include any alternate name to minimize false positives. The choice of whether to add the alternate names in each case was based on a preliminary analysis we performed on a small set of records to determine the ideal configuration for minimizing false positives and false negatives.

We used the developed knowledge base, along with a set of rule-based heuristics, to automatically extract, integrate, and normalize geospatial metadata from multiple fields in virus-related GenBank records (see Figure 3). We analyzed the following GenBank metadata fields of all virus sequences: *country*, *strain* and *isolate*. As mentioned earlier, the *country* field is the designated field in GenBank for storing information about the LOIH of virus sequences but additional geographic information may often be found in the *strain* field or the *isolate* field of GenBank records. In case of GenBank records pertaining to influenza viruses, the strain name of the virus sequence may also be recorded in the *organism* field of the record, which, therefore, presents another potential source of information for the LOIH of the virus, especially when the *strain* field is empty. However, many species of viruses, such as the Puumala virus, contain location mentions (in this case Puumala) within their species name which do not refer to the LOIH of the specific virus sequence. Therefore, to avoid the possibility of including erroneous locations, and for simplicity, our system only analyzed the *organism* field for influenza viruses.

For each GenBank record included in our database, our system first segmented the string in each pertinent field of the record based on simple delimiters, and considered each segment to be a possible candidate location. It then searched our developed knowledge base to find possible

matches for each candidate location. In case of overlapping locations, it chose the location with the greater number of tokens. For instance, if the content of the *country* field in a GenBank record is "Sierra Leone", our system would extract "Sierra Leone" [55] as a single location although GeoNames includes separate locations named "Sierra" [56] and "Leone" [44] respectively. To avoid false positives, our system discarded any candidate location which consisted of only three letters, unless it corresponded to a US state postal code (*e.g.* NY for New York). If no match was found, our system removed words such as "state", "county", "region", "east", "west" from the candidate location name and re-initiated the search. If still no match was found, our system applied a simple Lucene-based spell corrector to check for misspellings. The spell corrector first checked if a match could be found by inserting a space after each character in the query string to handle cases like "NewYork". If no match was found, it retrieved the top ten Lucene matches within two edit distance of the string if its length was greater than seven characters, and within one edit distance of the string if its length was greater than five characters but less than seven characters. Our system prioritized matches with the same phonetic representation as that of the query string (via the phonetic algorithm in Double Metaphone [57]) over those that did not have the same representation, and selected the top ranked match as the corrected spelling. It then integrated extracted location mentions to produce a coherent set of locations.

Our location integration algorithm functioned under the assumption that country-level locations are more likely to have been extracted correctly by our system than ADM1-level locations, which in turn are more likely to have been extracted correctly than locations less specific than ADM1. For instance, if our system extracted the locations "Grebe" and "Russia", it would disregard "Grebe" since, according to the GeoNames KB, there is no location called "Grebe" in Russia, and so it would assume that "Grebe" was extracted incorrectly. Similarly, if it extracted the locations "Grebe", "California' and "USA", it would once again disregard the location "Grebe", even though one exists in the state of Oregon in USA, since none can be found in California, USA. Before integrating any location more specific than ADM1-level, we ensured that it was contained within any ADM1-level or country-level location extracted from GenBank. We did not control the coherence between locations beyond the ADM1-level since we considered hierarchical data in GeoNames to be

adequately complete up to ADM1 level. For instance, we are confident that the GeoNames KB would include the parent country name and ADM1-level location name for all locations named “Grebe”. Therefore, our coherence checking process is more likely to lower false positives, than introduce false negatives. However, we are not as confident that GeoNames would correctly include the parent ADM2-level location and beyond for all locations, and so we chose not to check coherence beyond ADM1-level, to minimize the risk of missing valid locations. If multiple sets of coherent locations were found, our algorithm chose the set that provided more information. For instance, if our system extracted “Connecticut, USA”, and “Summit, New Jersey, USA”, it would choose the latter since that includes a larger set of coherent locations. After selecting a coherent set of locations, our algorithm outputted the most specific location in the set, along with its parent country and ADM1-level location, if available. For instance, if the selected set included “Chicago”, “Illinois”, and “USA”, our system would produce the integrated metadata “Chicago, Illinois, USA”. We included the parent country and ADM1-level locations to provide semantic context for our normalization algorithm, as we detail next.

When performing normalization, our system first searched our knowledge base of geographic locations to retrieve all possible GeoNames entries for the location, using the most comprehensive information available. For instance, if the integrated geographic metadata was "Chicago, Illinois, USA", a search was performed to retrieve all matches for "Chicago" in the state of Illinois in the country of USA and, thus, the locations “Illinois” and “USA” were used as semantic context by the normalization algorithm. Next, it narrowed the search results to the group of entries which possessed the least specific feature code (code in GeoNames denoting the type of the location, *e.g.*, country, state, city, etc.). For instance, in GeoNames, “Arizona” can be both a state in United States (with a feature code of ADM1) and a populated place in the state of Texas, United States (with a feature code of PPL) but our system would only select the former entry since it has a less specific feature code. This heuristic is based on the assumption that the less specific a location is, the more widely known it tends to be, and authors typically tend to refer to more widely known locations [23]. Our system then further narrowed down the search to the set of entries with the highest population. This heuristic is based on the assumption that geographic locations that have

higher populations tend to be referenced more often by authors [23], [32]. If the system still failed to uniquely identify the location, it randomly selected one of the possible entries. In case of records for which our system was unable to extract any location from any of the fields analyzed, the LOIH was listed as “Unknown” with a GeoNames ID of “-1”, and the latitude and longitude fields were populated with “0”.

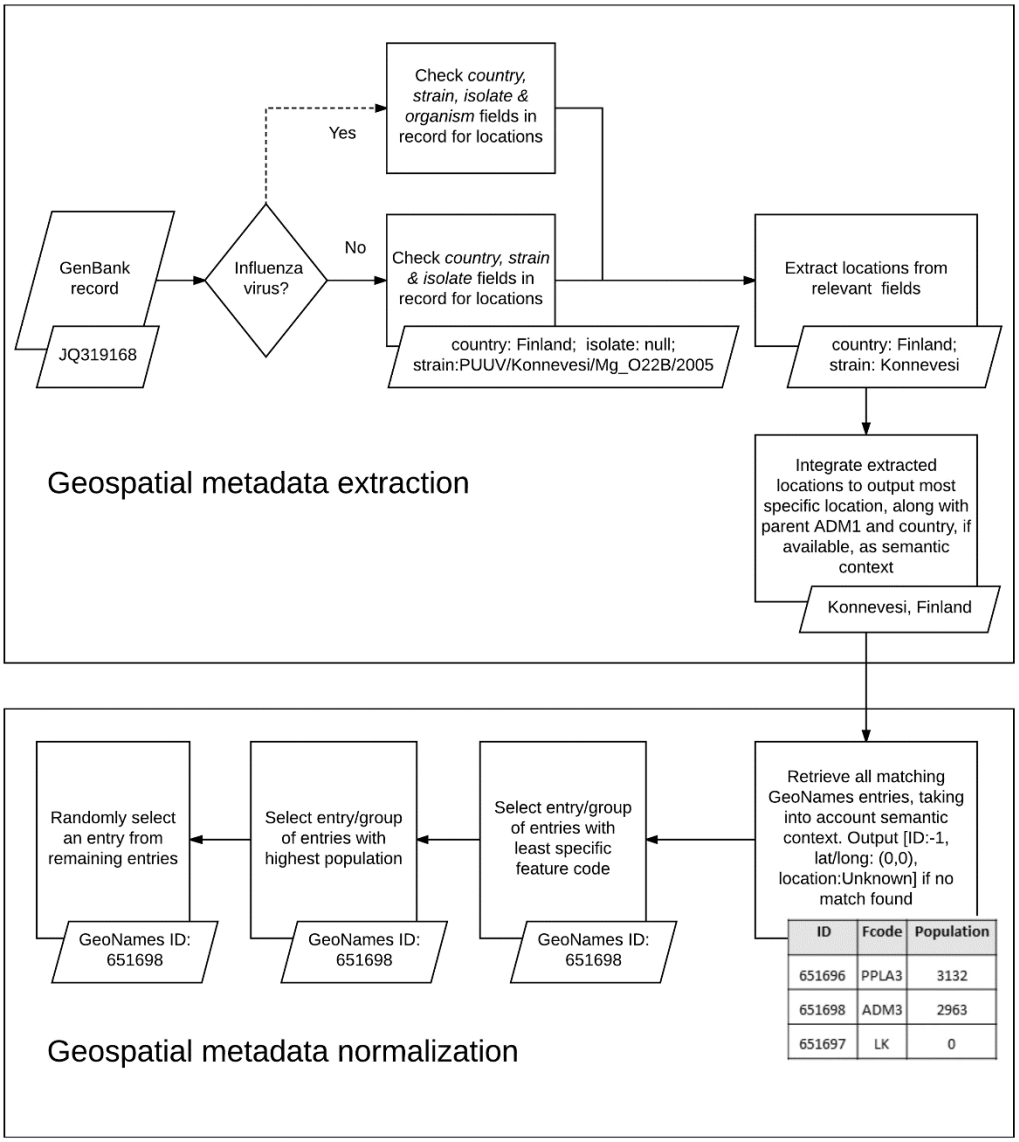


Figure 3. Geospatial metadata extraction and normalization.

As we outline in *Related Work*, our geospatial metadata extraction and normalization algorithm expands upon our prior work [23] in this area. However, we made a significant number of changes

to the pipeline to enhance its efficiency and accuracy, so that it may be more easily applied for the large-scale project undertaken here. Important updates include the following: 1) migration from MySQL database to Lucene index for storing the knowledge base of geographic locations in order to enable faster queries, 2) addition of rules to parse *strain* and *isolate* fields of non-influenza viruses, which tend to be less structured than those of influenza viruses, 3) addition of a simple spell-corrector to account for spelling errors in GenBank 4) addition of rules to allow the system to extract the most specific LOIH of the virus, instead of simply extracting any location more specific than ADM1-level (*e.g.* if the complete geospatial metadata was “Chicago, Cook County, Illinois, USA”, our prior algorithm may not extract “Chicago” since it would stop searching once it found “Cook County”), and 5) normalization to GeoNames IDs rather than latitude and longitude coordinates (this was a not a significant change with respect to implementation, but has important implications in supporting GenBank-related research by enabling cross-database queries).

3.4.3 Evaluation

To evaluate the accuracy of our normalization algorithms, we randomly selected 100 GenBank accession numbers among those included within our database, for manual annotation. Two annotators, whose biomedical specialties required them to work extensively with GenBank records pertaining to viruses, annotated and normalized the LOIH and infected host of each selected GenBank record. In both cases, the annotators used the GenBank website to acquire pertinent GenBank metadata for completing their annotation tasks. For each virus sequence, they tried to find the most comprehensive information available in GenBank concerning its LOIH and host name, using all fields present in GenBank, regardless of if our program used the field. In addition, in case of host metadata, our annotators also used domain knowledge to annotate host names even when they were not included in GenBank. For instance, they automatically assigned "host=human" to any HIV record. Also, the strain name of the influenza virus typically includes the name of its infected host, unless the host is human. Therefore, with the absence of any host metadata included in the strain field, it is reasonable to infer that the host was human for any non-laboratory strains.

After retrieving each relevant GenBank metadata, our annotators normalized it based on the selected knowledge base. For normalizing geospatial metadata, they searched each location in the

GeoNames website [12] to retrieve their corresponding GeoNames ID. Like our program, when multiple GenBank entries were available for a given location, they selected the one with the least specific feature code. For normalizing host names, our annotators used the NCBI Taxonomy website [58] to determine the Taxonomy ID of the host. In cases where they were unable to link a GenBank metadata to the selected knowledge base, they inserted '0' in the ID field.

Once the annotations were complete, we computed percentage agreement between our annotators for each annotation type to serve as a measure of inter-rater reliability. We chose to use percentage agreement, rather than Kappa statistic, because the number of possible categories in each annotation is over a million. The Kappa statistic is used to take into account agreement by chance [59]. However, given the number of possible categories in each annotation, the possibility of agreement by chance is negligible. Other studies in information retrieval have used f-score as a measure of agreement [23], [60]. However, in this study, each annotation simply involves entering a single value; therefore, the calculation of f-score would be redundant and a simple percentage agreement calculation is justified.

Once we completed the calculation of percentage agreement between our two annotators, a third annotator went through each case where they differed and selected the correct annotation to create our gold standard dataset. If it was unclear which annotation should be chosen, all annotators discussed the reason for the difference in annotation and mutually decided on one. Once the gold standard was created, we compared the annotations in the gold standard with the corresponding content in our database for measuring accuracy. In addition, to obtain a baseline performance measure for the host normalization task, we also computed the accuracy of the MetaMap [53] tool for this task on our gold standard dataset. When running MetaMap, we restricted sources to the NCBI Taxonomy vocabulary and retrieved the UMLS concept id with the highest score in each case. We then used the UMLS MRCONSO.RRF file to map the concept ids to their corresponding NCBI Taxonomy IDs. We applied the bias-corrected and accelerated (BCa) bootstrap method [61] with 10,000 iterations using the 'boot' package [62] in *R* to calculate the 95% confidence intervals (CI) for each accuracy value.

3.5 Results and Discussion

3.5.1 Database statistics

We provide key statistics pertaining to our database, which currently contains 2,244,971 GenBank records corresponding to 162,043 distinct virus organisms. We successfully mapped:

- The LOIH of 2,014,269 (89.7%) records to their respective GeoNames IDs by our LOIH normalization algorithm. Only 18,525 (0.8%) of these records originally had values in their "lat_lon" field.
- The infected hosts of 1,583,989 (70.6%) records to their respective NCBI Taxonomy ID by our host normalization algorithm. None of the GenBank records contained a formal link between the host field and an entry in the NCBI Taxonomy database.

3.5.2 Host Normalization Analysis

Rule-based methods are known to fail to capture the infinite variety of the human language, and consequently, our approach is expected to be imperfect. The host names of 29.4% of the GenBank records in our database were not normalized and were assigned the Taxonomy ID of 0. 27.4% were not normalized simply because the value in the *host* field was left empty. However, for 2.2%, 49,644 instances with instances repeated corresponding to 6803 unique instances, the *host* field contained a value but our rules failed to find the corresponding Taxonomy ID. We randomly selected 100 unique instances from this set and analyzed the reasons for the failure of the rules. For 41 instances, the presence of an abbreviation made the exact matching impossible, *e.g. C. tantalus* didn't match with *Chlorocebus tantalus* (ID: 60712) [63], and *tantalus* alone is not a concept in the taxonomy. For 35 instances, the *host* field contained the host name but also included additional information which was often not separated by delimiters from the host name, making the search difficult, *e.g. marine Heterobranchia species* where *Heterobranchia* is found in the taxonomy (ID: 216305) [64] but the presence of *marine* and *species* leads our algorithm to fail. The last 24 instances were not found in the taxonomy due to misspellings like *Lepus europeaus* for *Lepus europaeus* (ID:9983) [65], and missing entries in the Taxonomy database for alternative names of species such as *isard* for *Pyrenean chamois* (ID: 72545) [66], or species such as *Paradoxurus*

musangus. These reasons can also be found together in the same name of host making its normalization even more difficult. Further research is needed to design dedicated strategies to discover the reasons for the failure of the rules and finalize the normalization.

3.5.3 Geospatial Metadata Normalization Analysis

In case of geospatial metadata normalization, our system analyzed data from multiple fields in each GenBank record, and when running the pipeline, we recorded the number of locations extracted from each field. We found that our system extracted a total of 2,968,570 locations from the GenBank record fields analyzed. This count simply represents all unique locations extracted by our system per record, and includes duplicate locations, in cases where they were extracted from different records. For instance, given a sample of two records, one having the location "Chicago" in the *strain* field and "USA" in the *country* field, and the other having the location "USA" in both the *strain* and *country* fields, the total number of extracted locations, counted through this method, would be three. The percentage of the 2,968,570 locations that were extracted from the *country*, *strain*, o*rganism*, and *isolate* fields was 87.3% (2,594,402 locations), 17.3% (514,282 locations), 14.7% (434,931 locations), and 4.75% (141,064) respectively (the percentages do not add up to one since many of the locations were extracted from multiple fields *e.g.* if a location in a given record was collected from both the *strain* and *country* fields, it would be included in the count for both fields). Therefore, 12.7% of the locations were extracted from GenBank fields other than the *country* field. This indicates the importance of analyzing GenBank fields other than the *country* field for extracting geospatial metadata.

A comparison of the percentage of GenBank records in our database having missing values in the *country* field, with the percentage our algorithm failed to normalize, also illustrates the significance of integrating geospatial metadata from multiple GenBank record fields rather than only the *country* field. In 12.4% (278,350 records) of all GenBank records in our database, we did not find any data in the *country* field. However, we were unable to normalize the geospatial metadata of only 10.3% of the GenBank records in our database. This means that for at least 2.1% of the GenBank records, we added additional information from GenBank record fields other than the *country* field.

To obtain an estimate of the frequency with which our algorithm failed to extract geospatial data from the *country* field, we counted the number of records our algorithm failed to normalize despite the presence of geographic information in the *country* field. For a total of 2,310 records (0.1% of all records), representing 13 unique LOIH and 14 unique locations, the *country* field contained geospatial metadata which our algorithm was unable to extract. Of the 14 locations missed, seven were missed because we did not include the alternate names of all locations from GeoNames. For instance, GeoNames lists “British Guiana” as an alternate name for the main entry “Guyana”, but since we are not analyzing alternate names, our system failed to extract it. For four of the missed locations (*e.g.* “Kpokhankro”), we did not find any match in the GeoNames website when we manually searched for them. Therefore, the locations are most likely missing in GeoNames. Two of the locations were missed due to the presence of the word “the” before the geographic location mention *e.g.* “The Netherlands”. Although, as described in *Methods*, we used a list of stopwords to remove every GeoNames entry from our database which was an exact match for one of the stopwords in our list (such as “but”), we did not remove stop words from within GeoNames entries which were composed of multiple words. For instance, we did not remove the string “but” from within the GeoNames entry called “Ban Nong Yai But” [67], which represents a city/town in Thailand. Similarly, we did not remove stop words from within strings extracted from GenBank metadata such as “The Netherlands”. In case of locations for which our algorithm failed to find a match in GeoNames, it removed words such as “state”, “county”, “south” *etc.* and attempted to find a match again. However, stopwords such as “the” were not included in this list since their presence may possibly provide valuable context (as in the case of “But” in ““Ban Nong Yai But”). Moreover, our spell correction algorithm only searched for matches within 1 or 2 edit distance of the candidate string, depending on the string length. Addition of “the” represents an insertion of 4 additional characters (including space) and, therefore, our spell correction algorithm failed to find a match as well. The remaining location was missed due to the failure our algorithm to correctly identify the abbreviation “USSR” standing for the former Union of Soviet Socialist Republic, now dissolved.

3.5.3.1 Annotation statistics

Our gold standard annotation dataset includes 100 GenBank records with 64 distinct LOIH and 20 distinct host names. The percent agreement between our annotators for host and LOIH normalization was 95 and 83 respectively (Table 1). Differences in host annotation resulted from either of the two annotators missing a host name present in GenBank, erroneously adding a host name not present, not selecting the most specific host name available (*e.g.* deer instead of roe deer) or not annotating the host name of a record with missing host metadata even when it could be inferred based on the virus organism. In case of geospatial metadata annotation, our annotators annotated the same location in seven of the 17 instances where their final ID annotation differed. However, they disambiguated the locations differently. The remaining differences arose from missed locations.

Task	Inter-rater Agreement (%)	System Accuracy (%)
Host metadata normalization	95	70
LOIH metadata normalization	83	87

Table 1: Inter-rater agreement and accuracy of normalization tasks based on manually created gold standard of 100 GenBank records

3.5.4 Accuracy Statistics

We found that the accuracy of our normalization algorithms for host and geospatial metadata to be 70% (95% CI [0.60-0.77]) and 87% (95% CI [0.78-0.92]) respectively when evaluated on the manually annotated gold standard (Table 1). The baseline performance for host normalization using MetaMap was found to be 63% (95% CI [0.52-0.71]).

Of the 30 errors in host normalization, 28 were from lack of domain knowledge, 27 of which were specifically the result of the program not knowing that certain viruses affected only a single species of organism. One error resulted from the inability of the system to extract the host 'bar-headed goose' and in case of the remaining error, the program correctly extracted the host name but incorrectly normalized it to the ID of its parent organism.

MetaMap's accuracy was found to be 7% lower than that of our system. As expected, MetaMap missed all host names where domain knowledge was required. In addition, in many cases it mapped the entities to higher level concepts than what was annotated by our annotators. For instance, MetaMap normalized the host "duck" to the concept id corresponding to the genus "Anas"[68] while our annotators normalized it to the taxonomy id corresponding to the specific species "Anas platyrhynchos" [69].

Of the 13 errors in LOIH geospatial metadata normalization, eight were due to disambiguation errors (same string representation of locations but different GeoNames IDs), three were due to missed locations, and the remaining two were due to the detection of locations not annotated by our annotators. The disambiguation errors resulted from the inability of our algorithm to choose the correct location based on exact string match. For instance, based on our annotation guidelines, the annotators normalized the location "Ningbo, Zhejiang, China" to the GeoNames ID "1799395" [70] which corresponds to the second order administrative division (ADM2) named "Ningbo Shi" in GeoNames. Since "Ningbo Shi" is not an exact match for "Ningbo", our program incorrectly normalized it to the GeoNames ID "1799397" [71] instead, which corresponds to the capital city of Ningbo Shi, and is named "Ningbo" in GeoNames. Among the missed location errors, one was a result of GeoNames including a different spell variant of the location, which our system was not able to recognize. The remaining locations were missed because they were annotated based on the *title* field in GenBank (a field containing the title of a publication linked to the record) but our program does not extract metadata from the *title* field. Both of the locations extracted by our program but not annotated in our gold standard were valid locations. We chose to not include one of them in our gold standard because it was too ambiguous and it was not possible to correctly normalize it based on available information. The other was most likely missed by both of our annotators.

3.6 Conclusion

In this study, we developed an automated framework for extracting and normalizing two different types of GenBank metadata which are widely used in different domains of biomedical research. We applied our framework to retrieve the host and geospatial metadata of over two million

GenBank records pertaining to viruses, and link them to the NCBI Taxonomy database and the GeoNames database respectively. We have made the database including the normalized metadata publicly available to allow researchers to easily integrate them within their works and help accelerate biomedical discovery. In addition, we also created a manually annotated gold standard dataset consisting of 100 randomly selected GenBank records for evaluating the normalization algorithms. The percent agreement between our annotators was over 80% for the annotation of the two GenBank metadata types, which is adequately high. It was higher (95%) for host annotation than for geospatial metadata annotation (83%), illustrating the latter to be the more challenging of the two tasks when performed manually.

When evaluated on the gold standard set, our host and LOIH normalization algorithms achieved accuracies of 70% (95% CI [0.60-0.77]) and 87% (95% CI [0.78-0.92]) respectively. The majority of the errors in host normalization resulted from lack of domain knowledge, indicating the need to incorporate additional rules within our system to account for cases where a virus organism may only infect a single type of host organism. However, our current lack of such rules should not in any way reduce the applicability of our released dataset, since researchers are more likely to utilize it for GenBank records where the host name is not definitively known based on the nature of the virus organism. Our system correctly normalized the host name in all but two records, where specific domain knowledge was not required. In case of geospatial metadata normalization, the accuracy of our system was in fact higher than the inter-rater agreement calculated for its annotation. The systematic nature of our algorithm made it more suitable for this difficult task which requires extensive efforts when done manually.

Our gold standard dataset is relatively small and, since it was randomly selected, it often included duplicates of the most common hosts or geographic locations included in GenBank, leading to an even lower number of distinct metadata annotations. This is especially true for host metadata annotation, since the infected host in most records included within our dataset was ‘human’. Therefore, our measured performance may not be reflective of the average performance of the algorithms in other datasets.

In addition to evaluating our normalization algorithms on the gold standard dataset, we also performed supplementary analysis to investigate the completeness of the normalized host and geospatial metadata in our database. Our analysis showed that our normalization algorithms could normalize nearly all GenBank records for which the *host* and *country* fields were not empty. Although this does not necessarily mean that the extraction and normalization of metadata was performed correctly in each case, it nevertheless provides a simple measure of our system's ability to extract metadata whenever available. In addition, our analysis also revealed the importance of including geospatial metadata from different GenBank record fields rather than only the *country* field.

Although various large-scale efforts are currently being made by NCBI to facilitate curation of rich and consistent GenBank metadata, we believe that the framework and database presented in this manuscript would continue to remain highly useful to researchers, both in the present time and in the near future. Manual annotation of millions of GenBank records could take years and is an expensive process. In contrast, our normalization algorithms take only a few seconds to process each record and could help considerably accelerate this process. Moreover, it also provides a standardized method for disambiguating metadata and may even help correct some errors made by humans. We have already applied our framework to create a comprehensive database including the normalized host and geospatial metadata of over two million GenBank records, which is easily accessible online. Our database has the potential to support a wide range of large-scale analyses involving viruses and would greatly benefit researchers working with virus GenBank records. Furthermore, by providing a thorough description and analysis of the geospatial and species metadata normalization methods we developed through our project, we hope to assist researchers working with similar normalization problems in any field.

We have made the source code for our framework available through github and it can be easily extended to other pathogens as well. The *country* and *host* field of all entries in GenBank are similarly formatted, regardless of which organism it pertains to. Therefore, it should be possible to use our framework, as it is, for extracting metadata from these fields for any pathogen. However, unlike viruses, most pathogens do not contain additional information pertaining to their location of

collection in the other GenBank record fields analyzed by our algorithm, such as the *strain* field and *organism* field. Therefore, the inclusion of such fields would be unnecessary for other pathogens.

As future work, we plan to evaluate our existing algorithms on larger datasets, and work on improving their accuracy by including additional features such as a more sophisticated spell corrector. We also intend to use information extraction techniques to extract additional information about the locations of the infected host often mentioned in the unstructured texts of the *notes* and *comments* metadata fields. Our future work would also include exploring additional resources containing species information, such as Interagency Taxonomic Information System [72], Encyclopedia of Life [73], and Catalogue of Life [74], for host name normalization instead of relying solely on the NCBI Taxonomy database, which has missing entries for the alternative names of some organisms. In addition, we intend to modify our host normalization algorithm so that it is capable of recognizing varying degrees of taxonomy hierarchy, thereby allowing normalization to different levels of the taxonomy tree based on user needs. Addition of this feature would facilitate the normalization of host names which cannot be mapped to a single organism, since they could instead be mapped to higher nodes in the taxonomy tree, and would be highly useful in case of viruses which may live in different host organisms. Further important steps we plan to take through future work include developing normalization algorithms for other metadata in GenBank and extending our normalization algorithms to other non-virus organisms.

3.7 Funding

Research reported in this publication was supported by the National Library of Medicine (NLM) of the National Institutes of Health (NIH) under award number R01LM012080 to MS and National Institute of Allergy and Infectious Diseases (NIAID) of the NIH under award number R01AI117011 to MS and GG. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

3.8 References

[1] D. A. Benson *et al.*, “GenBank.,” *Nucleic Acids Res.*, vol. 41, pp. D36-42, Jan. 2013.

[2] R. Leinonen *et al.*, “The European nucleotide archive,” *Nucleic Acids Res.*, vol. 39, no.

SUPPL. 1, 2011.

[3] J. Mashima *et al.*, “DNA data bank of Japan (DDBJ) progress report,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D51–D57, 2016.

[4] “GenBank and WGS Statistics.” [Online]. Available: https://www.ncbi.nlm.nih.gov/genbank/statistics/. [Accessed: 28-Apr-2017].

[5] C. R. Howard and N. F. Fletcher, “Emerging virus diseases: can we ever expect the unexpected?,” *Emerg. Microbes Infect.*, vol. 1, no. 12, p. e46, Dec. 2012.

[6] J. P. Messina *et al.*, “Global spread of dengue virus types: mapping the 70 year history.,” *Trends Microbiol.*, vol. 22, no. 3, pp. 138–46, Mar. 2014.

[7] M. Scotch *et al.*, “Phylogeography of influenza A H5N1 clade 2.2.1.1 in Egypt.,” *BMC Genomics*, vol. 14, p. 871, Dec. 2013.

[8] “Mumps virus M, F, SH, HN, L genes for matrix protein, fusion protein, - Nucleotide - NCBI.” [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/AB618040. [Accessed: 01-May-2017].

[9] “Hepatitis A virus genomic RNA, nearly complete genome, isolate: HAJIH- - Nucleotide - NCBI.” [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/AB618529. [Accessed: 01-May-2017].

[10] “Cowpox virus partial A36R ortholog gene for p43-50 protein, strain VIS - Nucleotide - NCBI.” [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/AJ312308. [Accessed: 01-May-2017].

[11] “Influenza A virus (A/Toronto/T5294/2009(H1N1)) segment 1 sequence - Nucleotide - NCBI.” [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/CY045959. [Accessed: 01-May-2017].

[12] “GeoNames.” [Online]. Available: http://www.geonames.org/. [Accessed: 05-Sep-2013].

[13] P. Gratton, S. Marta, G. Bocksberger, M. Winter, E. Trucchi, and H. Kühl, “A world of sequences: can we use georeferenced nucleotide databases for a robust automated phylogeography?,” *J. Biogeogr.*, vol. 44, no. 2, pp. 475–486, 2017.

[14] S. Federhen, “The NCBI Taxonomy database.,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D136-43, Jan. 2012.

[15] “Home - BioSample - NCBI.” [Online]. Available: https://www.ncbi.nlm.nih.gov/biosample. [Accessed: 16-Sep-2017].

[16] T. Barrett *et al.*, “BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata.,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D57-63, Jan. 2012.

[17] N. A. O’Leary *et al.*, “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D733-45, Jan. 2016.

[18] “RefSeq: NCBI Reference Sequence Database.” [Online]. Available: https://www.ncbi.nlm.nih.gov/refseq/. [Accessed: 16-Sep-2017].

[19] “Virus Variation.” [Online]. Available: https://www.ncbi.nlm.nih.gov/genome/viruses/variation/. [Accessed: 16-Sep-2017].

[20] E. L. Hatcher *et al.*, “Virus Variation Resource - improved response to emergent viral outbreaks.,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D482–D490, Jan. 2017.

[21] “Viral Genomes.” [Online]. Available: https://www.ncbi.nlm.nih.gov/genome/viruses/. [Accessed: 16-Sep-2017].

[22] J. R. Brister, D. Ako-Adjei, Y. Bao, and O. Blinkova, “NCBI viral genomes resource.,” *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D571–7, Jan. 2015.

[23] T. Tahsin *et al.*, “A high-precision rule-based extraction system for expanding geospatial metadata in GenBank records,” *J. Am. Med. Informatics Assoc.*, vol. 23, no. 5, pp. 934–941, Sep. 2016.

[24] E. W. Bloomquist, P. Lemey, and M. A. Suchard, “Three roads diverged? Routes to phylogeographic inference,” *Trends Ecol. Evol.*, vol. 25, no. 11, pp. 626–632, 2010.

[25] I. N. Sarkar, “Leveraging biomedical ontologies and annotation services to organize microbiome data from Mammalian hosts.,” *AMIA Annu. Symp. Proc.*, vol. 2010, pp. 717–721, 2010.

[26] M. A. Musen *et al.*, “The National Center for Biomedical Ontology.,” *J. Am. Med. Inform. Assoc.*, vol. 19, no. 2, pp. 190–5, 2012.

[27] E. S. Chen and I. N. Sarkar, “Towards Structuring Unstructured GenBank Metadata for Enhancing Comparative Biological Studies.,” *AMIA Jt. Summits Transl. Sci. Proc. AMIA Summit Transl. Sci.*, vol. 2011, pp. 6–10, Jan. 2011.

[28] L. Sinclair *et al.*, “Seqenv : linking sequences to environments through text mining,” *PeerJ*, vol. 4, p. e2690, Dec. 2016.

[29] P. L. Buttigieg, N. Morrison, B. Smith, C. J. Mungall, S. E. Lewis, and ENVO Consortium, “The environment ontology: contextualising biological and biomedical entities.,” *J. Biomed. Semantics*, vol. 4, no. 1, p. 43, Dec. 2013.

[30] D. Sharma, P. Priyadarshini, and S. Vrati, “Unraveling the web of viroinformatics: computational tools and databases in virus research.,” *J. Virol.*, vol. 89, no. 3, pp. 1489–501, Feb. 2015.

[31] J. Tamames and V. de Lorenzo, “EnvMine: a text-mining system for the automatic extraction of contextual information.,” *BMC Bioinformatics*, vol. 11, p. 294, Jan. 2010.

[32] J. L. Leidner, “Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding,” *SIGIR Forum*, vol. 41, no. 2, pp. 124–126, Dec. 2007.

[33] D. Weissenbacher *et al.*, “Knowledge-driven geospatial location resolution for phylogeographic models of virus migration.,” *Bioinformatics*, vol. 31, no. 12, pp. i348–i356, Jun. 2015.

[34] M. Gerner and G. Nenadic, “Named Entity Recognition and Normalization of Species, LINNAEUS,” in *Encyclopedia of Systems Biology*, New York, NY: Springer New York, 2013, pp. 1489–1492.

[35] L. Tari, S. Anwar, S. Liang, J. Hakenberg, and C. Baral, “Synthesis of pharmacokinetic

pathways through knowledge acquisition and automated reasoning.," *Pac. Symp. Biocomput.*, pp. 465–76, Jan. 2010.

[36] W. Shen, J. Wang, P. Luo, and M. Wang, "LIEGE: Link Entities in Web Lists with Knowledge Base," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, 2012, p. 1424.

[37] G. Limaye, S. Sarawagi, and S. Chakrabarti, "Annotating and searching web tables using entities, types and relationships," *Proc. VLDB Endow.*, vol. 3, no. 1–2, pp. 1338–1347, 2010.

[38] N. Limsopatham and N. Collier, "Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation," *Proc. 54th Annu. Meet. Assoc. Comput. Linguist. (Volume 1 Long Pap.*, pp. 1014–1023, 2016.

[39] Q. Zhu, R. R. Freimuth, J. Pathak, and C. G. Chute, "PharmGKB Drug Data Normalization with NDF-RT.," *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.*, vol. 2013, p. 180, 2013.

[40] M. Bada, "Mapping of biomedical text to concepts of lexicons, terminologies, and ontologies," *Methods Mol. Biol.*, vol. 1159, pp. 33–45, 2014.

[41] J. L. Bono *et al.*, "Phylogeny of shiga toxin-producing escherichia coli o157 isolated from cattle and clinically ill humans," *Mol. Biol. Evol.*, vol. 29, no. 8, pp. 2047–2062, 2012.

[42] "Homo sapiens." [Online]. Available: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=9606. [Accessed: 01-May-2017].

[43] "Federal Republic of Germany, Germany - A PCLI 2921044." [Online]. Available: http://www.geonames.org/2921044/federal-republic-of-germany.html. [Accessed: 01-May-2017].

[44] "Leone, American Samoa - P PPL 5881347." [Online]. Available: http://www.geonames.org/5881347/leone.html. [Accessed: 15-May-2017].

[45] W. Shen, J. Wang, and J. Han, "Entity linking with a knowledge base: Issues, techniques, and solutions," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 443–460, 2015.

[46] "Murine norovirus strain MuNoVIT1, complete genome - Nucleotide - NCBI." [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/KR349276. [Accessed: 05-May-2017].

[47] "Mus pahari." [Online]. Available: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=10093. [Accessed: 01-May-2017].

[48] "Mus musculus." [Online]. Available: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=10090. [Accessed: 01-May-2017].

[49] "Mus spretus." [Online]. Available: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=10096. [Accessed: 01-May-2017].

[50] "Paris, United States - P PPLA2 4717560." [Online]. Available: http://www.geonames.org/4717560/paris.html. [Accessed: 16-Sep-2017].

[51] “Paris, France - P PPLC 2988507.” [Online]. Available: http://www.geonames.org/2988507/paris.html. [Accessed: 16-Sep-2017].

[52] “Nanovirus-like particle isolate WOK1, complete sequence - Nucleotide - NCBI.” [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/KT390491. [Accessed: 01-May-2017].

[53] “MetaMap - A Tool For Recognizing UMLS Concepts in Text.” [Online]. Available: https://metamap.nlm.nih.gov/. [Accessed: 16-Sep-2017].

[54] “Country List ISO 3166 Codes Latitude Longitude | Socrata.” [Online]. Available: https://opendata.socrata.com/dataset/Country-List-ISO-3166-Codes-Latitude-Longitude/mnkm-8ram. [Accessed: 18-Jun-2014].

[55] “Republic of Sierra Leone, Sierra Leone - A PCLI 2403846.” [Online]. Available: http://www.geonames.org/2403846/republic-of-sierra-leone.html. [Accessed: 15-May-2017].

[56] “Sierra County, United States - A ADM2 5395582.” [Online]. Available: http://www.geonames.org/5395582/sierra-county.html. [Accessed: 15-May-2017].

[57] “DoubleMetaphone (Apache Commons Codec 1.10 API).” [Online]. Available: https://commons.apache.org/proper/commons-codec/apidocs/org/apache/commons/codec/language/DoubleMetaphone.html. [Accessed: 05-May-2017].

[58] “Home - Taxonomy - NCBI.” [Online]. Available: https://www.ncbi.nlm.nih.gov/taxonomy. [Accessed: 28-Apr-2017].

[59] M. L. McHugh, “Interrater reliability: the kappa statistic.,” *Biochem. medica*, vol. 22, no. 3, pp. 276–82, 2012.

[60] G. Hripcsak and A. S. Rothschild, “Agreement, the f-measure, and reliability in information retrieval.,” *J. Am. Med. Inform. Assoc.*, vol. 12, no. 3, pp. 296–8, Jan. 2005.

[61] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman and Hall., 1993.

[62] “Package ‘boot.’” [Online]. Available: https://cran.r-project.org/web/packages/boot/boot.pdf.

[63] “Taxonomy browser (Chlorocebus tantalus).” [Online]. Available: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=60712. [Accessed: 31-May-2017].

[64] “Taxonomy browser (Heterobranchia).” [Online]. Available: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=216305&lvl=3&lin=f&keep=1&srchmode=1&unlock. [Accessed: 31-May-2017].

[65] “Taxonomy browser (Lepus europaeus).” [Online]. Available: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9983. [Accessed: 31-May-2017].

[66] “Taxonomy browser (Rupicapra pyrenaica).” [Online]. Available: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=72545&lvl=3&lin=f&keep=1&srchmode=1&unlock. [Accessed: 31-May-2017].

[67] “Ban Nong Yai But, Thailand - P PPL 8370083.” [Online]. Available: http://www.geonames.org/8370083/ban-nong-yai-but.html. [Accessed: 16-Sep-2017].

[68] “Anas.” [Online]. Available: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=8835. [Accessed: 03-Oct-2017].

[69] “Anas platyrhynchos.” [Online]. Available: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=8839&lvl=3&lin=f&keep=1&srchmode=1&unlock. [Accessed: 03-Oct-2017].

[70] “Ningbo Shi, China - A ADM2 1799395.” [Online]. Available: http://www.geonames.org/1799395/ningbo-shi.html. [Accessed: 01-May-2017].

[71] “Ningbo, China - P PPLA2 1799397.” [Online]. Available: http://www.geonames.org/1799397/ningbo.html. [Accessed: 01-May-2017].

[72] “Background Information.” [Online]. Available: https://www.itis.gov/info.html. [Accessed: 16-Sep-2017].

[73] “Encyclopedia of Life.” [Online]. Available: http://eol.org/. [Accessed: 16-Sep-2017].

[74] “Home | Catalogue of Life.” [Online]. Available: http://www.catalogueoflife.org/. [Accessed: 16-Sep-2017].

4 GEOBOOST: ACCELERATING RESEARCH INVOLVING THE GEOSPATIAL METADATA OF VIRUS GENBANK RECORDS

Authors: Tasnia Tahsin, Davy Weissenbacher, Karen O'Connor, Arjun Magge, Matthew Scotch, and Graciela Gonzalez-Hernandez

4.1 Abstract

Summary: GeoBoost is a command-line software package developed to address sparse or incomplete metadata in GenBank sequence records that relate to the location of the infected host (LOIH) of viruses. Given a set of GenBank accession numbers corresponding to virus GenBank records, GeoBoost extracts, integrates and normalizes geographic information reflecting the LOIH of the viruses using integrated information from GenBank metadata and related full-text publications. In addition, to facilitate probabilistic geospatial modeling, GeoBoost assigns probability scores for each possible LOIH.

Availability and implementation: Binaries and resources required for running GeoBoost are packed into a single zipped file and freely available for download at https://tinyurl.com/geoboost. A video tutorial is included to help users quickly and easily install and run the software. The software is implemented in Java 1.8 and supported on MS Windows and Linux platforms.

4.2 Introduction

Locations of infected hosts (LOIH) are critical pieces of metadata required for exploring the spread and evolutionary dynamics of pathogens such as viruses. This information is often retrieved from GenBank, a public database of nucleotide sequences which is maintained by the National Center of Biotechnology Information (NCBI) [1]. Researchers have used the geospatial metadata in virus GenBank records for a wide range of public health studies. For instance, the LOIH of viruses based on GenBank metadata have been used to map the global spread of viruses [2], investigate the environmental predictors of virus diffusion [3], and trace the origin of infectious disease outbreaks [4].

Currently, the extraction of the LOIH of viruses is performed manually and requires a significant investment of time and effort. The designated field for storing the LOIH of viruses in GenBank

records is the *country* field, which despite its name, may contain geospatial metadata of varying degrees of specificity. For instance, the *country* field of the GenBank record with accession no. CY058987 (https://www.ncbi.nlm.nih.gov/nuccore/CY058987) contains the province-level metadata "China: Hubei". Due to the nature of virus nomenclature, additional geospatial metadata may often be found in the *strain* and *isolate* fields of GenBank records. In the aforementioned GenBank record, the *strain* field contains the location "Wuhan" embedded in the strain name "A/Wuhan/390/2005". Thus, researchers often need to manually integrate locations from different GenBank record fields to retrieve the most specific spatial resolution. Other times, the geospatial metadata available in GenBank is not sufficient for a given study [5]. For instance, a researcher modeling the spread of the rabies virus within an US state would likely need at least the county-level LOIH of all virus samples, but GenBank may not contain such precise information. In such cases, researchers often search full-text publications linked to GenBank for more specific information. Moreover, depending on the type of study, they may also need to normalize each extracted LOIH to its latitude/longitude coordinates (*e.g.* for continuous phylogeography) or to a standardized string representation (*e.g.* for discrete phylogeography). The ambiguity of locations (e.g. Paris can be in France or Texas, USA) makes this task especially challenging.

We present GeoBoost, a knowledge-driven framework for automatically extracting, integrating and normalizing the LOIH of viruses from GenBank records and related full-text articles. It builds upon our prior work in this area [6], including additional features to enhance its usability and performance. To the best of our knowledge, this is the only publicly accessible software available for this task. Related work has been performed to extract, and often normalize, different forms of GenBank metadata [7]–[10]. For instance, the Tempus et Locus (TeL) software was developed to extract GenBank sequences containing the date of collection and/or location of sampling of the sequence [10]. However, GeoBoost is the only software we know of which attempts to enhance existing geographic metadata in GenBank by extracting and integrating geographic information from related full-text publications.

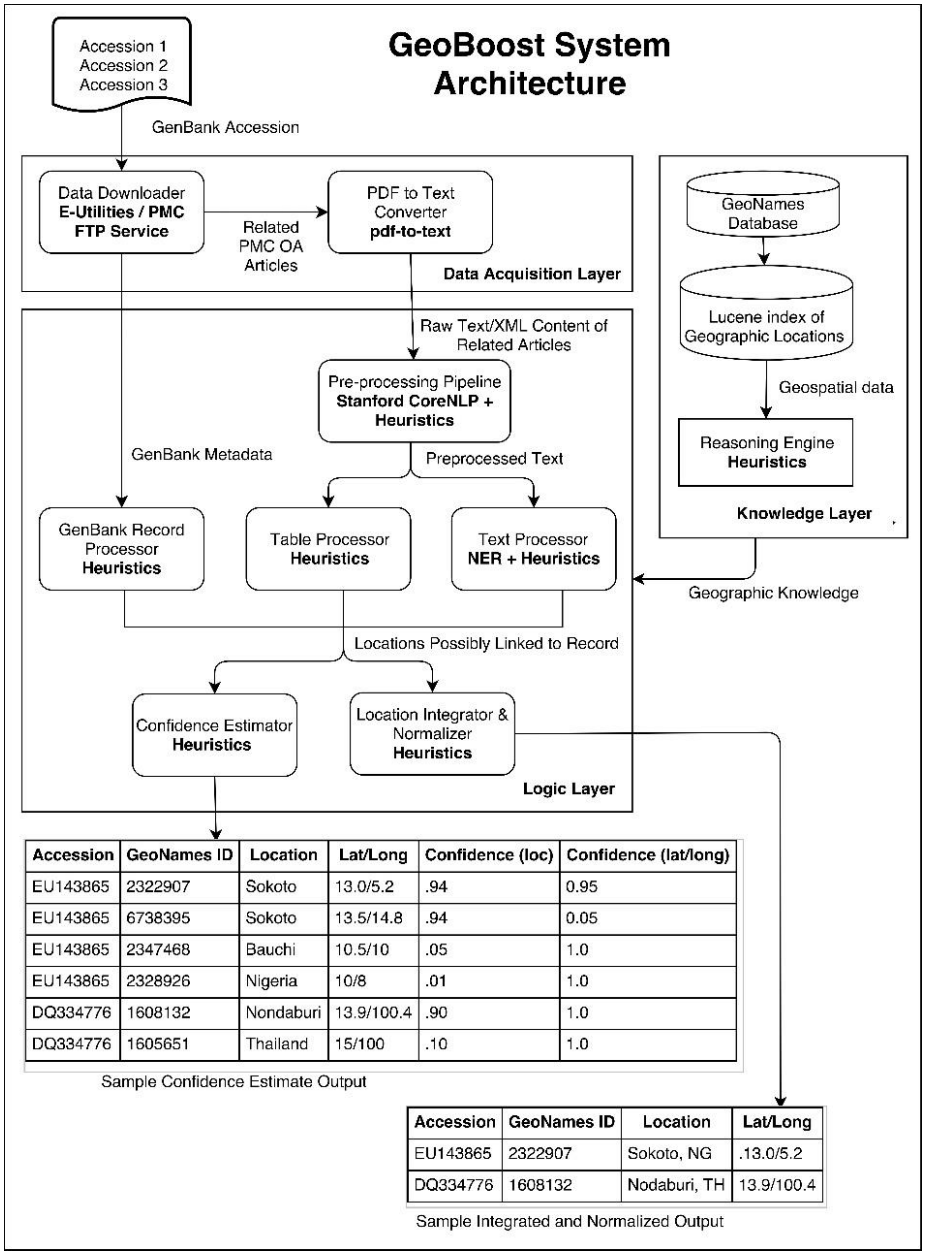
In addition to outputting the most specific and most probable LOIH of a virus, GeoBoost also assigns a probability score to each possible LOIH of the virus based on its specificity (e.g. Phoenix

is more specific than Arizona) and likelihood of being correct. Researchers can then use these scores for selection of geographic locations to build precise models of virus spread. Additionally, if no further location information can be found, it will also indicate this "failure", saving researchers additional time in ruling out locations. GeoBoost is easy to install and run, and could accelerate bioinformatics research that incorporates the LOIH of viruses.

4.3 Materials and Methods

GeoBoost (see Figure 1) is a knowledge-driven framework and central to its function is a knowledge base (KB) of geographic locations. Our KB is primarily based on the GeoNames.org database, which contains geospatial data for over 10 million geographic locations. Given a list of GenBank accession numbers, GeoBoost uses the Entrez Programming Utilities [11] to download relevant metadata from the corresponding GenBank records. It also downloads all PubMed Central (PMC) Open Access articles linked to each record in both PDF and XML format, if available, and converts the PDF files to text files using the pdf-to-text software (http://www.foolabs.com/xpdf/home.html). It then uses knowledge-driven heuristics to extract and integrate geospatial metadata from relevant fields in each GenBank record, until a user-provided sufficiency criterion is satisfied. For instance, if the sufficiency criterion is "ADM1" (*i.e.* states or provinces of a country), GeoBoost will stop searching once it finds ADM1-level or more specific geographic information. If GeoBoost fails to find sufficient geospatial metadata for a record even after analyzing all relevant fields, it proceeds to search the free-text and tabular content of linked articles. If GeoBoost is not given a sufficiency criterion, it searches all available sources for the most specific LOIH.

When searching the tabular content of an article associated with a GenBank record, GeoBoost uses simple rules to analyze the structure and content of each table and extract possible links between GenBank records and geographic locations. When searching the free-text content of a related article, GeoBoost applies a Named Entity Recognition system [12] for detecting geographic location mentions in text and uses rule-based heuristics to determine which of the extracted locations are more likely to be linked to the record.



Accession	GeoNames ID	Location	Lat/Long	Confidence (loc)	Confidence (lat/long)
EU143865	2322907	Sokoto	13.0/5.2	.94	0.95
EU143865	6738395	Sokoto	13.5/14.8	.94	0.05
EU143865	2347468	Bauchi	10.5/10	.05	1.0
EU143865	2328926	Nigeria	10/8	.01	1.0
DQ334776	1608132	Nondaburi	13.9/100.4	.90	1.0
DQ334776	1605651	Thailand	15/100	.10	1.0

Accession	GeoNames ID	Location	Lat/Long
EU143865	2322907	Sokoto, NG	.13.0/5.2
DQ334776	1608132	Nodaburi, TH	13.9/100.4

Figure 1 GeoBoost System Architecture. Given a user-provided list of GenBank accession numbers corresponding to viruses, the Logic Layer uses the geographic knowledge provided by the Knowledge Layer, and the GenBank metadata and PMC OA articles downloaded by the Data Acquisition layer to output: 1) the most probable, integrated, normalized location of infected host (LOIH) of each virus (integrated and normalized output), and 2) the probability scores of each possible LOIH of each virus (confidence estimate output)

After extracting locations from all possible sources, GeoBoost integrates and normalizes them, and outputs the most probable and most specific LOIH of each virus (P location | GenBank record). For instance, if it extracted “USA” from the GenBank record, “Paris” from the free-text content of a related article, and “Texas” from the tabular content of a related article, it would output “Paris, Texas, USA”, given there is more evidence for the later than for “Paris, France”. GeoBoost normalizes each LOIH to its corresponding GeoNames ID and latitude/longitude coordinates based on rule-based heuristics. It also assigns probability scores to each possible LOIH of a virus using a complex set of heuristics that assigns higher scores to more accurate and more specific locations. The probability score assignment process is performed in two stages. In the first stage, GeoBoost assigns probability scores to every location extracted by the pipeline from either the GenBank record or linked article. In the next stage, it assigns probability scores to all possible latitude/longitude pairs associated with each candidate location in GeoNames. Probability scores assigned in both steps add up to 1.0.

To estimate the performance of GeoBoost, we used two different manually annotated sets of GenBank records, created through our prior work [6]. The first set (*Flu*) included 5728 GenBank records corresponding to influenza viruses. We annotated the LOIH of the viruses in this set based on information in the GenBank records and 60 full-text PMC articles linked to these records. The second set (*Non-Flu*) included 100 GenBank records corresponding to six different non-influenza viruses. We annotated the LOIH of the viruses in this set based on information in the GenBank records and 10 full-text PMC articles linked to these records. For accuracy, we calculated the percentage of the records for which the top ranked latitude/longitude coordinates outputted by GeoBoost was within 50 miles of the manually annotated latitude/longitude coordinates. We also calculated the time taken by GeoBoost to process each record using a JVM heap size of 1GB, a download speed of ~100 Mbps, and the Windows 10 Operating System. We measured the accuracy and time taken by GeoBoost for each dataset under two different settings: 1) when configured to download and extract information from related PMC OA articles along with GenBank metadata (default configuration), 2) when configured to use GenBank metadata only. This allowed us to

assess the added benefit of extracting and integrating information from related articles, in addition to using GenBank metadata

4.4 Results

In Table 1 we show the results of our evaluation. Under default configuration, GeoBoost had a high level of accuracy for both test sets (81% for Flu and 80% for Non-Flu), and took less than 5 seconds to process each record (1.59s for Flu and 4.65s for Non-Flu). When configured to exclude information from related PMC OA articles, GeoBoost's accuracy fell by 11% for the Flu set and 25% for the Non-Flu set, demonstrating the value of extracting and integrating additional information from related articles.

Test Set	Accuracy (%)		Time per record (s)	
	Including PMC OA	Excluding PMC OA	Including PMC OA	Excluding PMC OA
Flu_Set	81	70	1.59	1.01
Non-Flu_Set	80	55	4.65	1.40
Average	**80.5**	**62.5**	**3.12**	**1.21**

Table 1. Performance evaluation of GeoBoost relative to manually annotated gold standard

4.5 Acknowledgements

We would like to thank Robert Rivera, Rachel Beard, Mari Firago and Dr. Garrick Wallstrom for their contributions. RR, RB and MF annotated data for the test set. RR and GW contributed to confidence estimate process. RR reviewed the paper.

4.6 Funding

Research reported here was supported by the NIAID of the NIH under Award Number R01 AI117011 to GG and MS. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Conflict of Interest: None declared.

4.7 References

[1] D. A. Benson *et al.*, “GenBank.,” *Nucleic Acids Res.*, vol. 41, pp. D36-42, Jan. 2013.

[2] J. P. Messina *et al.*, “Global spread of dengue virus types: mapping the 70 year history.,” *Trends Microbiol.*, vol. 22, no. 3, pp. 138–46, Mar. 2014.

[3] D. Magee, R. Beard, M. A. Suchard, P. Lemey, and M. Scotch, “Combining phylogeography and spatial epidemiology to uncover predictors of H5N1 influenza A virus diffusion.,” *Arch. Virol.*, vol. 160, no. 1, pp. 215–24, Jan. 2015.

[4] R. G. Wallace and W. M. Fitch, “Influenza A H5N1 immigration is filtered out at some international borders.,” *PLoS One*, vol. 3, no. 2, p. e1697, Jan. 2008.

[5] M. Scotch *et al.*, “Enhancing phylogeography by improving geographical information from GenBank.,” *J. Biomed. Inform.*, vol. 44 Suppl 1, pp. S44-7, Dec. 2011.

[6] T. Tahsin *et al.*, “A high-precision rule-based extraction system for expanding geospatial metadata in GenBank records,” *J. Am. Med. Informatics Assoc.*, vol. 23, no. 5, pp. 934–941, 2016.

[7] I. N. Sarkar, “Leveraging biomedical ontologies and annotation services to organize microbiome data from Mammalian hosts.,” *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2010, pp. 717–21, Nov. 2010.

[8] E. S. Chen and I. N. Sarkar, “Towards Structuring Unstructured GenBank Metadata for Enhancing Comparative Biological Studies.,” *AMIA Jt. Summits Transl. Sci. Proc. AMIA Summit Transl. Sci.*, vol. 2011, pp. 6–10, Jan. 2011.

[9] P. Gratton, S. Marta, G. Bocksberger, M. Winter, E. Trucchi, and H. Kühl, “A world of sequences: can we use georeferenced nucleotide databases for a robust automated phylogeography?,” *J. Biogeogr.*, vol. 44, no. 2, pp. 475–486, 2017.

[10] A. R. Carter and D. Gatherer, “Tempus et Locus: a tool for extracting precisely dated viral sequences from GenBank, and its application to the phylogenetics of primate erythroparvovirus 1 (B19V),” *bioRxiv*, Jan. 2016.

[11] E. Sayers, “E-utilities Quick Start.” National Center for Biotechnology Information (US), 09-Aug-2013.

[12] D. Weissenbacher *et al.*, “Knowledge-driven geospatial location resolution for phylogeographic models of virus migration.,” *Bioinformatics*, vol. 31, no. 12, pp. i348–i356, Jun. 2015.

4.8 Appendix A. GeoBoost Architecture Description

The GeoBoost system architecture may be broadly divided into three layers: Knowledge Layer, Data Acquisition Layer and Logic Layer. Below, we provide detailed descriptions of each of these layers.

4.8.1 Knowledge Layer

The knowledge layer is primarily based on the GeoNames database downloaded from http://download.geonames.org/export/dump/ on June 13, 2017. GeoNames contains over 10 million place names in the world. In addition, each entry in GeoNames also includes features such as the population, latitude/longitude and type (e.g. state, country, city etc.) of the location. The type of the location is represented using a field called feature code. GeoNames has over 645 feature codes to indicate the type of a location. Moreover, GeoNames also contains information about the hierarchical structure of locations by including the parent country and administrative division of every entry.

Two major challenges associated with using GeoNames for detecting place names in text are: 1) it contains many place names (such as But, David etc.) which are used to denote English stop words or other named entities, such as names of people, more frequently than names of places, and 2) it is a very large database and traditional methods for representing dictionaries in Named Entity Recognition (NER) systems require too much memory. For instance, the use of Patricia trees for representing GeoNames required over 9 GB of memory.

To address the first challenge, we filtered out place names in GeoNames based on manually created lists of English stop words such as “but”, frequently used words in the biomedical domain such as “gene”, frequently used names of people etc. To address the second challenge, we stored the GeoNames database as a Lucene Index which enabled very fast queries.

In addition to the Lucene index of geographic locations, the Knowledge Layer also includes a Reasoning Engine for enabling spatial reasoning such as checking whether two place names are coherent or not. Coherency between place names is checked by exploiting information about the hierarchical structure of locations in GeoNames. For instance, the place names “Grebe” and

"China" would be considered incoherent since there is no location called Grebe in the country China while the place names "Guangdong" and "China" would be considered coherent since Guangdong is a province in China. The Reasoning Engine also enables coherence checking between a place name and a set of coherent place names. For instance, "Texas, USA" and "Paris" would be considered coherent since Texas, USA has a city called Paris. However, "Texas, USA" and "Seattle" would be considered incoherent since there is no location called "Seattle" in Texas, USA.

4.8.2 Data Acquisition Layer

Given a list of accession numbers, the Data Acquisition Layer downloads the following GenBank record fields using the Entrez Programming Utilities (E-Utilities): *definition, strain, isolate, organism, country, date of collection, host* and *gene.* The downloaded metadata is stored as a tab-separated text file. In addition, the Data Acquisition Layer also uses the E-Utilities to retrieve PMCIDs for all PMC articles linked to each GenBank accession. It then uses the PMC FTP Service to download the subset of the linked PMC articles which are available through PMC Open Access (OA). PMC OA articles are available in XML and PDF format. The pdf-to-text software from XPDF (https://www.xpdfreader.com/download.html) is used to convert the PDF files to raw text files.

4.8.2.1 Logic Layer

The Logic Layer uses the geographic knowledge provided by the Knowledge Layer, and the GenBank metadata and PMC OA articles downloaded by the Data Acquisition layer to output: 1) the most probable, integrated, normalized location of infected host (LOIH) of each virus (integrated and normalized output), and 2) the probability scores of each possible LOIH of each virus (confidence estimate output). The individual steps performed in this process are described below:

1. Extract coherent set of locations from GenBank metadata: The first step in the Logic Layer involves using the Record Processor to extract and integrate a coherent set of locations from the downloaded GenBank metadata until a user-provided sufficiency criterion is satisfied. The GenBank metadata fields analyzed includes the *country*, *strain,* and *isolate* fields. In case of influenza viruses (the *definition* field is used to determine whether or not a GenBank record

represents an influenza virus), the Record Processor also analyzes the *organism* field which often contains the complete strain name of the virus even when the *strain* field is missing. The *organism* field is not analyzed for other viruses since their organism names often contain place names which do not refer to the LOIH of the virus (e.g. the organism name “Puumala virus” contains the place name “Puumala” which does not necessarily represent the LOIH of the virus).

To extract locations, the Record Processor first segments the string in each GenBank record field analyzed based on a set of delimiters, and considers each segment to be a possible candidate location. It then uses the Knowledge Layer to search for possible matches for each candidate location. If no match is found, the Record Processor applies a domain-specific spell corrector to check for misspellings.

When integrating locations, the Record Processor once again relies heavily on the Knowledge Layer. Location integration is performed through coherency checking. For instance, if the *strain* field of a record contained the place name “Guangdong” and the *country* field contained the place name “China”, the Record Processor would extract “Guangdong, China”, since Guangdong is a province in China. However, if the *strain* field contained the place name “Grebe” and the *country* field contained the place name “China”, the Record Processor would only output “China” since there is no location called “Grebe” in “China” (hence they are incoherent). In case of inconsistent locations, preference is given to locations extracted from the *country* field over those extracted from other fields since the former is more straightforward and likely to result in fewer errors. Preference is also given to less specific place names over more specific place names e.g. if the *country* field of a record contained the two place names “China” and “Grebe”, the country name “China” would be selected instead of the place name “Grebe”, which can be a populated place in Bosnia and Herzegovina, a lake in Canada or a park in USA (all instances are less specific than country). This is because less specific place names are typically more widely known than more specific place names, and therefore referenced more often.

After extracting and integrating locations from each field, the Record Processor checks if a user-provided sufficiency criterion is satisfied before analyzing remaining fields. For instance, if the user only needs country-level or state-level locations for his/her study and the Record Processor

extracts “Guangdong, China” from the *country* field, it would not analyze any other field in the GenBank record since Guangdong is a province in China, and therefore the extracted metadata already satisfies the sufficiency criterion. If the sufficiency criterion is set to “none”, GeoBoost extracts all available geographic metadata for each record. The Record Processor analyzes the GenBank record fields in the following order: *country, strain, isolate, organism* (note: *organism* field is analyzed for the influenza virus only).

2. Extract and link locations from related PMC OA articles: If the geographic metadata extracted and integrated from all analyzed fields in GenBank is found to be insufficient, the Logic Layer searches for more specific information in related PMC OA articles. To perform this task, it first pre-processes the raw text files (obtained through conversion of PDF files to text files) and XML files of the related PMC OA articles. The raw text files are used for extracting information from the unstructured content of the articles. The XML files are used for extracting information from the tabular content of the articles (the corresponding text files are not used for this purpose since the tables do not retain their structure after PDF-to-text conversion). When preprocessing the XML files, the XML structure is exploited to extract all relevant tables in each file. A table is considered to be relevant if one of its columns contains geographic information (location column) while another contains either the strain name, date of collection or name of host of GenBank sequences (GenBank metadata column). Table headers are used to determine the type of information contained in each column. For instance, a table column is considered to possess geographic information if its header contains the word “location” or any of its synonyms such as "province", "state", "origin", "country", "place", "region", "district", "area" etc. The Table Processor compares the data in the GenBank metadata column of each row in each preprocessed relevant table with the corresponding data in the GenBank record being analyzed until a match is found. Places names present in the location column of the matching row are then linked to the record. The Knowledge Layer is used to detect place names in the location column.

The raw text files obtained through PDF-to-text conversion are preprocessed using a Natural Language Processing (NLP) Pipeline based on the Stanford CoreNLP package for tagging sentences, tokens, and parts-of-speech. A Named Entity Recognition (NER) system, which utilizes

the Lucene index in the Knowledge Layer as its dictionary of place names, is then applied to detect place names in text. The Text Processor analyzes every sentence corresponding to each place name detected by the NER to see if it matches any one of a set of manually defined rules for linking GenBank record to geographic location mentions in text. The rules were derived through analysis of 27 articles in a preliminary study. Given a sentence containing a place name, the Text Processor first checks if the sentence matches the very general "Location Pattern", given in Table 2. If it matches this pattern, then the Text Processor proceeds to check if the sentence matches any of the rules in the Very High, High, Mid, Low or Very Low rule levels (see Table 3), in the order specified. For instance, if the GenBank record corresponding to a H1N1 virus indicates that the virus was isolated in the year 2009 from ducks, and the location under consideration is China, then the sentence "We collected H1N1 viruses from ducks in China in the year 2009" will match one of the rules in the "Very High" rule level, while the sentence "We collected H1N1 viruses from China" will only match one of the rules in the "Low" rule level. The level of a rule is based on its discriminatory power for establishing GenBank record-location linkages.

3. Assign Confidence Estimates to Extracted Locations: After extracting locations from GenBank metadata and the tabular and textual content of related articles, the Logic Layer uses the Confidence Estimator to generate confidence estimates to every GeoNames entry corresponding to the locations extracted by Record Processor, Table Processor and Text Processor. The Confidence Estimator applies a heuristic algorithm to perform this task in two steps. In the first step, it assigns confidence estimates to every place name extracted (including those extracted by the NER which did not match any of the linkage rules used by the Text Processor); next, it assigns confidence estimates to all possible GeoNames entry associated with each place name. We do not directly assign confidence estimates to all possible GenBank record-GeoNames entry pairs since that would unfairly penalize locations that are more ambiguous. That is, some locations in GeoNames have over a hundred candidates and this might result in a diluted, very low confidence for each candidate for these locations if we were to treat them all "equally". Instead, we process these locations through our linkage analysis, and end up with a set of "reasonable" linkages with a

specific confidence assigned to them. Thus, the algorithm for our linkage analysis, which serves as the first phase in our confidence assignment process, is based on the following principles:

- Every location extracted should be assigned some confidence estimate, however small
- The confidence estimates assigned should reflect both the likelihood of the location being correct as well as the specificity of the location since more specific and more probable locations are likely to produce more precise geospatial models of virus spread
- Separate confidence estimates should be assigned to overlapping locations since there is a chance one is correct while the other is not. For instance, if we extract two locations, Paris and France, we should assign separate confidence estimates to each because there is a chance Paris is correct but France is not (it may be from Paris, Texas, USA) and there is also a chance that France is correct but Paris is not (it may be from a different location in France). Moreover, since the aim of this system is to increase the specificity of insufficient GenBank records, we would also want the more specific location to be given a higher weight.
- The confidence estimate assigned to a location mention should depend on the precision of its source of extraction. Table 3 lists the precision of the major sources of locations used by our system. The precision of each source was calculated by evaluating it against a small manually annotated test set. Since a few of the sources had a precision of 1, we subtracted a small value of 0.02 from the measured precision of every source when using it for confidence estimate generation of locations in order to satisfy condition 1.
- Locations extracted from the geospatial metadata of the GenBank records with greater specificity should be given priority over those with lower specificity by being assigned a higher confidence estimate. For instance, if a record contains both country-level locations and ADM1-level locations consistent with the country (i.e. present inside the country), then the ADM1-level location should be prioritized since it is of greater interest to the user.
- Locations which are extracted from either the textual or tabular contents of the paper should be assigned higher confidence estimates if they increase the specificity of existing locations in the GenBank Record.

- Locations which are extracted from either the textual or tabular contents of the paper should be penalized if they are inconsistent with the locations already present in the GenBank record. An extracted location is inconsistent if it cannot be present in the same country and ADM1-level location mentioned in the GenBank record.
- The confidence estimate generated for locations extracted from the unstructured textual content of related PMC OA articles should take into account the precision of the rule-level used to extract the location. We calculated the precision and recall of each rule level using a manually annotated set of GenBank records (see Table 4).
- The (normalized) individual confidence estimates should add up to 1

The final implementation of this stage is largely dependent on the number of unique sources analyzed by our system for determining the location of collection and uses a complex set of heuristics.

In the second phase of the confidence assignment process, the Confidence Estimator assigns confidence estimates to every GeoNames entry corresponding to every location mention. If a place name is associated with only a single entry in GeoNames, then the corresponding GeoNames entry is assigned 100% confidence. The second phase of the confidence assignment process is based on the following principles:

- A candidate which is consistent with existing geospatial metadata should be assigned higher confidence over inconsistent candidates. For instance, the place name Paris can have a candidate in both USA and France but if the geospatial metadata says 'France', the GeoNames entry associated with Paris in France should be assigned the highest confidence estimates.
- A less specific candidate should be assigned higher confidence over more specific candidates. For instance, Texas can refer to either the state of Texas in USA or a populated place in New York, USA. The former is assigned higher confidence estimate since it is less specific.
- A more populated candidate should be assigned higher confidence estimate over a less populated candidate. This is a popular, widely used heuristic for disambiguation

which assumes that locations with higher populations tend to be more popular and therefore appear more frequently in textual sources of information.

4. *Integrate and normalize linked locations:* The Data Integrator and Normalizer applies a heuristic algorithm to integrate and normalize locations extracted from the GenBank record and the unstructured and tabular content of related PMC OA articles. Like the Record Processor, it checks the coherence between locations to integrate them. Locations inconsistent with geographic metadata extracted from GenBank are discarded. Higher preference is given to locations linked to a GenBank record based on the tabular content of related PMC OA articles than to locations linked to a GenBank record based on the unstructured textural content of related PMC OA articles since the former has higher precision. The Data Integrator attempts to output the most specific and most likely LOIH of each virus based on a set of heuristics and includes ADM1-level and country-level information for the location, when available, for semantic context e.g. if the system extracts “USA” from the GenBank metadata, “Paris” from the tabular content of a related article, and “Texas" from the unstructured textual content of a related article, it would output “Paris, Texas, USA”. The Normalizer attempts to disambiguate the location extracted by the Data Integrator. To perform this task, the Normalizer first retrieves all GeoNames entries corresponding to the integrated location produced by the Data Integrator. For instance, if the integrated location is “Cambridge, USA”, it would retrieve all GeoNames entries corresponding to the place name “Cambridge” which has a country code of “US”. The Normalizer then sorts the retrieved GeoNames entries based on their feature codes and chooses the group of entries belonging to the least specific feature codes. For instance, as mentioned earlier, Texas can refer to either the state of Texas in USA or a populated place in New York, USA. The Normalizer would prioritize the GeoNames entry corresponding to the state of Texas in USA. Lastly, the Normalizer sorts the group of entries selected in the previous step based on their population, and outputs the set with the highest population.

Location Pattern	**".* (in|from|at) .*"+location+".*";**
Isolated pattern	".* (isolated|collected) .* (in|from) .*"+location+".*" ".* we .* (collect|isolate) .* (in|from) .*"+location+".*" ".* (isolation|collection) .* (in|from) .*"+location+".*"
We Rule	".* we used .*" ".* in this study .*" ".* our study .*" ".* we examined .*" ".* we studied .*" ".* current study .*"

Table 2: General patterns applied by different levels of rules used in the Text Processor for linking place names in unstructured textual content of related PMCOA articles to GenBank records

Rule level	Rules present	Precision, Recall
Very High	allPresent(strain, date, host) && sentence.contains(strain) && sentence.contains(year) && sentence.contains(host) && hasIsolatedPattern(sentence, location)==1	1.0, 0.024
	allPresent(accession) && sentence.contains(accession) && sentence.contains("isolated")	
	allPresent(date, host, virusName) && sentence.contains(year) && sentence.contains(host) && sentence.contains(virusName) && (sentence.contains("strain") || sentence.contains("isolate")) && hasIsolatedPattern(sentence, location)==1)	
High	allPresent(strain, date) && sentence.contains(strain) && sentence.contains(year) && hasIsolatedPattern(sentence, location)==1	0.995, 0.047
	allPresent(strain, host) && sentence.contains(strain) && sentence.contains(host) && hasIsolatedPattern(sentence, location)==1	
	allPresent(accession) && sentence.contains(accession))	
	allPresent(date, virusName) && sentence.contains(year) && sentence.contains(virusName) && (sentence.contains("strain") || sentence.contains("isolate")) && hasIsolatedPattern(sentence, location)==1	

	allPresent(host, virusName) && sentence.contains(host) && sentence.contains(virusName) && (sentence.contains("strain") || sentence.contains("isolate")) && hasIsolatedPattern(sentence, location)==1)	
Mid	allPresent(strain, date) && sentence.contains(strain) && sentence.contains(year)	1.0, 0.21
	allPresent(strain, host) && sentence.contains(strain) && sentence.contains(host)	
	allPresent(date, host) && sentence.contains(year) && sentence.contains(host)&&weRule(sentence)==1)	
	allPresent(date, virusName)==1 && sentence.contains(year) && sentence.contains(virusName) && (sentence.contains("strain") || sentence.contains("isolate"))	
	allPresent(host, virusName)==1 && sentence.contains(host) && sentence.contains(virusName) && (sentence.contains("strain") || sentence.contains("isolate")	
Low	allPresent(date, host)==1 && sentence.contains(year) && sentence.contains(host))	0.884, 0.278
	sentence.contains(year) && date.length()>0) || (sentence.contains(host) && host.length()>0)) && weRule(sentence)==1	
	allPresent(virusName) && sentence.contains(virusName) && weRule(sentence)==1	
	allPresent(strain) && sentence.contains(strain)	
	allPresent(virusName) && sentence.contains(virusName) && (sentence.contains("strain") || sentence.contains("isolate")) && hasIsolatedPattern(sentence, location)==1	
	sentence.contains(virusName) && ((sentence.contains(host) && host.length()>0) || (sentence.contains(year) && date.length()>0)))	
Very Low	sentence.contains(year) && date.length()>0 sentence.contains(host) && host.length()>0 sentence.contains(virusName) && virusName.length()>0	0.643, 0.541

	weRule(sentence)==1 hasIsolatedPattern(sentence, location)==1	

Table 3: Rules used by the Text Processor for linking place names in unstructured textual content of related PMCOA articles to GenBank records. The rules are organized into different levels based on their discriminatory power. Please note that the AllPresent(String[] args) method returns true if the GenBank record has non-null entries for the argument fields. E.g. AllPresent(strain, date, host) means that the strain field, date field and host fields are not null or blank in the record.

Source	Precision	Confidence (precision -0.02)	Recall
Record Processor	0.954	0.934	0.996
Table Processor (consistent subset only)	1.0	0.980	0.838
Text Processor Linkage Extraction Algorithm (consistent subset only)	0.847	0.827	0.799
Very High	1.0	0.980	0.033
High	0.999	0.979	0.111
Mid	1.0	0.980	0.129
Low	0.962	0.942	0.537
Very Low	0.836	0.816	0.717
Text Processor NER (consistent subset only)	0.599	0.579	0.904
Overall Extraction	0.876	0.856	0.978

Table 4: Confidence of Each Source of Extraction

5 DISCUSSION

The results achieved through the works presented in this dissertation illustrate knowledge-driven methods to be highly effective for geo-referencing virus GenBank records. Chapter 2 introduced the basic framework of a system for extracting, integrating, and normalizing the LOIH of viruses based on information present in GenBank records and related full-text articles. It found the knowledge-driven methods used in this framework to have an f-score of 0.894 for linking GenBank records of influenza viruses to the specific latitude and longitude coordinates of their LOIH. To test the generalizability of this framework to other viruses, we performed a crude evaluation using a smaller test set incorporating GenBank records pertaining to six other viruses, and the system was found to have an accuracy of 0.75 for identifying the correct LOIH of viruses in this set.

The work described in Chapter 3 enhanced the system's Record Location Extractor module, which is the system component responsible for extracting geographic metadata from different fields in virus GenBank records. It then applied this enhanced version of the module to geo-reference all virus GenBank records available at the time of the study. The resulting database, consisting of over two million virus GenBank records, was made freely accessible to the public both as a downloadable file and as a navigable website. When evaluated on a manually annotated test set, the database was found to have a high accuracy of 87% for linking virus GenBank records to the correct GeoNames ID of their sampling location. The system was modified to link GenBank records to specific GeoNames IDs instead of only mapping them to the latitude and longitude coordinates of their sampling sites (as in Chapter 2) because this enabled a more uniform normalization, linking two well-known databases, and the resulting unique IDs could be directly used in discrete phylogeography or easily translated to geo-coordinates for use in continuous phylogeography. In addition to describing the development and evaluation of this normalized database, Chapter 3 also presented a thorough analysis of the database demonstrating the significance of integrating geographic metadata from different fields in GenBank records instead of simply using the designated "country" field for storing the LOIH of the viruses.

Chapter 4 presented GeoBoost, the final integrated system for geo-referencing GenBank records which was made publicly available for download along with a video tutorial providing step-

by-step instructions to facilitate its use among researchers. GeoBoost had a more structured architecture than the system introduced in Chapter 2, and automated tasks, such as data download, which had to be performed separately in the earlier system. In addition, it introduced a new Confidence Estimator module which assigned confidence estimates to every possible GeoNames entry corresponding to the locations extracted by GeoBoost from GenBank records and the textual and tabular content of related articles. These confidence estimates served as a measure of the specificity of the extracted location (e.g. Phoenix is more specific than Arizona) as well as its likelihood of being correct, enabling researchers to use these estimates for choosing the correct geographic locations when building precise models of virus spread. When evaluated on the same test sets used in Chapter 2, GeoBoost was found to have an accuracy of 81% and 80% for the influenza set and the non-influenza set respectively.

Please note that the results are not directly comparable between Chapter 2 and Chapter 4 because we used different evaluation criteria. The system described in Chapter 2 did not assign confidence estimates to different locations. Instead, it used a heuristic algorithm to extract and disambiguate all locations that it deemed to have a high likelihood of being linked to the record, and, as described in Chapter 2, when estimating its performance on the influenza dataset, the f-score of the system was computed by counting the number of false positives, true positives and false negatives it produced. On the non-influenza test set, a simpler estimate of system accuracy was manually calculated by considering each record to be correctly processed or incorrectly processed based on whether the location extracted by the system, prior to being mapped to its latitude/longitude coordinates, was the same as the annotated location. We did not measure the disambiguation performance of the system for the non-influenza test. In contrast to the earlier version of the system, GeoBoost utilized the Confidence Estimator module to assign confidence estimates to the GeoNames IDs of all possible linked locations. As a measure of its performance, we calculated Geoboost's accuracy in Chapter 4 by considering a given record to be correctly processed if the latitude/longitude coordinates of the GeoNames entry ranked highest by the system was within 50 miles of the annotated latitude/longitude coordinates in the gold standard (the gold standard contained the latitude/longitude coordinates of the sampling site of each GenBank

record without including the GeoNames ID of the selected entry and so directly matching GeoNames IDs, as we did in Chapter 3, was not possible without additional annotation).

In addition to demonstrating the ability of GeoBoost to accurately and efficiently geo-reference virus GenBank records, Chapter 4 also established the importance of incorporating information from related full-text articles when attempting to extract the most precise LOIH of viruses. GeoBoost's accuracy fell by 11% and 25% for the influenza dataset and the non-influenza dataset respectively when configured to exclude information from related full-text articles and use the GenBank data alone. Being, to the best of my knowledge, the only system to-date capable of integrating geographic information pertaining to the LOIH of viruses from different fields in GenBank records as well as the textual and tabular content of related articles, this illustrates the value of using GeoBoost in the phylogeography research community where, as shown by Magee *et al*. [23], the incorporation of precise geographic information in the phylogeographic models could be very useful in understanding the predictors of viral spread.

Despite its recent release, GeoBoost has been used in several research studies analyzing virus migration patterns. For instance, Scotch *et al* [34] utilized GeoBoost in our study investigating the effect of incorporating sampling uncertainty of the LOIH of taxa in virus phylogeography. Also, Fisk [35] utilized GeoBoost in her study analyzing the global transmission patterns of the Respiratory Syncytial Virus (RSV). As a unique system addressing a significant challenge, GeoBoost has the potential to accelerate many other research studies involving the extraction of the LOIH of viruses. In addition, the knowledge-driven methods used in GeoBoost could be easily extended to non-virus taxa as well to help analyze their evolution and geographic distribution through time.

The works described in this dissertation have also helped guide several research studies involving the more general task of toponym resolution (detection and disambiguation) in full-text articles linked to GenBank records. For instance, one of my contributions in [57] involved the development of the "metadata heuristic" for toponym disambiguation, which extended the knowledge-driven method for disambiguating the geographic metadata of GenBank records, as described in Chapter 2, to assist the disambiguation of every geographic location mentioned in full-text publications linked to GenBank, thereby significantly broadening its applicability in biomedical

research. Magge *et al* [49] further extended this algorithm by adopting the approach taken by Tamames *et al.* [51] of including parent locations, if present in contiguous text following the place name mention, when performing the disambiguation, and reported a 3% increase in accuracy. However, as discussed in Chapter 1, it is not clear whether this increase in the toponym disambiguation accuracy in full-text articles linked to GenBank would help boost GeoBoost's accuracy. GeoBoost's primary objective is to geo-reference GenBank records and the extraction of geographic locations from related articles is one of the many tasks it performs to help it complete this objective. Currently, GeoBoost only performs disambiguation of locations after extracting and integrating information from the different textual sources it analyzes. For instance, if it extracts "USA" from the country field of a GenBank record, "Wisconsin" from the free-text content of the related article based on a high-precision sentence such as "All samples were collected from the state of Wisconsin.", and "Madison" from the tabular content of the article, it will consider the integrated location "Madison, Wisconsin, USA" to be a highly likely candidate for the sampling location of the record. Given the number of sources GeoBoost analyzes for performing this integration, it needs to apply a complex set of knowledge-driven heuristics and, currently, these heuristics do not take into account how individual locations within related articles were disambiguated. After producing the integrated locations, GeoBoost searches GeoNames for possible matches and ranks the GeoName IDs associated with each match. For instance, "Madison" in Wisconsin, USA can refer to a park in Milwaukee or the city of Madison, and GeoBoost will consider the latter to be the more likely candidate as it is a more widely referenced place. Therefore, given the way GeoBoost is currently designed, the toponym disambiguation accuracy in full-text articles does not affect its disambiguation accuracy for the LOIH of viruses and it is not clear whether incorporating the additional information will yield significant benefit.

The simple, primarily dictionary-based algorithm used in GeoBoost for toponym detection in full-text articles linked to GenBank was found have an f-score of 0.698 in [50] but recent work in this area, involving state-of-the-art neural network models, have reported significantly higher f-scores with Magge *et al*. [49] achieving an f-score of 0.94 for this task. However, the precise effect of the increase in performance of these toponym detection algorithms in GeoBoost is still unclear

as they have not been integrated into GeoBoost yet. For instance, these toponym detection methods achieved significantly higher precision than GeoBoost's NER but since GeoBoost already uses many knowledge-driven heuristics for filtering out false positive locations prior to linking place names extracted from related articles to GenBank records, to account for the low precision of its NER component, it is not clear whether this increase in precision will lead to a notable increase in GeoBoost's accuracy. Moreover, the rule-based NER in GeoBoost continues to maintain the highest recall for the task of toponym resolution (detection and disambiguation) in full-text articles, even though higher recall has been reported for toponym detection alone [49]. This illustrates that the rule-based NER method utilized in GeoBoost is still the most effective method for minimizing any false negatives for which GeoNames contains a matching entry. For instance, a sophisticated ML method might be able to correctly label "Monofeya" in the sentence "The study was performed in Monofeya" as a toponym. However, GeoNames currently does not have a match for Monofeya, which is precisely why the dictionary-based NER is unable to extract it. Therefore, GeoBoost will be unable to integrate it with existing geographic metadata in GenBank records related to the article, and despite an increase in recall of the NER system a corresponding increase in accuracy will not be seen in GeoBoost. To be able to harness the increase in NER recall, a spell-correction method will need to be applied to any toponyms extracted by the NER which does not have a match in GeoNames. Although GeoBoost utilizes such an algorithm to extract names of places mentioned in the GenBank record fields which do not produce a direct match in GeoNames, at its current state it does not use spell correction on the results of the NER system. Future work might involve adding this feature in GeoBoost in addition to enhanced NER systems for toponym detection.

The NER component in GeoBoost is not the only module that would benefit from additional improvements. It is important to address the limitations of other modules in the system as well. One of the limitations of the Record Location Extractor module in GeoBoost is its inability to extract individual locations from compounded strings. For instance, the strain field of GenBank record JQ714202 [30] contains "A/Tianjinheping/SWL313/2009" which includes two separate locations from China – the Tianjin province and the Heping district of the Tianjin province. Although GeoBoost utilizes different spell-correction heuristics to enhance its location extraction performance, which

enables it to extract, for instance, “New York” from “NewYorrk” and “Qalyubiya” from the spell variant “Qalyoubeya”, it is still not capable of extracting the two separate locations “Tianjin” and “Heping” from the compounded string “Tianjinheping”. Also, in many GenBank records, airport codes in the country, strain or isolate fields are used to indicate the LOIH of the virus e.g. the isolate field of Genbank record LC071961 [58] contains “BKK-TH-171”, where BKK is the airport code for Bangkok. The Record Location Extractor module in GeoBoost can expand acronyms of US states (e.g. CA for California) but is currently not capable of correctly mapping airport abbreviations.

On updating the Record Location Extractor to enable it to extract locations from compounded strings and map airport codes to locations, it achieved an accuracy of 0.95 on a manually annotated set of 5901 virus-related GenBank records for linking each GenBank record to a GeoNames ID based on geographic metadata present across the different fields in the records. However, the updated Record Location Extractor was not included in the version of GeoBoost released to the public because processing compounded strings for toponym extraction significantly slowed down GeoBoost. Future work may involve including a more efficient algorithm for extracting locations from compounded strings and evaluating the precise effect of the addition of these features more thoroughly.

The extent to which airport codes are used in GenBank to indicate the LOIH of viruses also needs further investigation. In the manually annotated test set, place names were only extracted from obvious airport codes, *e.g.* Bangkok from BKK. The updated Record Location Extractor was able to use it for less well-known airport codes as well but to adequately verify whether the increase in false positives this might possibly lead to is compensated by the increase in true positives and decrease in false negatives, our annotation schema will need to be changed to include more in-depth look into these abbreviations. For instance, the strain name of the virus represented by GenBank record JX262205 [59] is “A/India/GWL01/2011”. Since GWL is the airport code for the city of Gwalior in India, the updated Record Location Extractor mapped this record to “Gwalior, India”. However, the annotators only mapped the record to the country of India based on record metadata alone and so this instance was counted as an error of the system during the evaluation process since the system result did not match the annotated gold standard. A review of the related

article [60] reveals that the sampling site of this record is highly likely to be Gwalior, India. Therefore, this specific case was most probably wrongly labeled as an error during the evaluation process.

Currently the entire GeoBoost infrastructure relies primarily on handwritten knowledge-driven heuristics which requires a lot of time and effort to maintain and update. The incorporation of machine learning methods could allow the system to self-learn and adapt to different domains with minimal changes to its framework. The Text Linker module in GeoBoost, which is responsible for linking place names extracted from related full-text publications to GenBank records, is probably most likely to benefit from such a shift. Currently, the Text Linker module uses very simple patterns for linking GenBank records to locations detected in related articles. Although, as described in Chapter 2, it still managed to achieve an f-score of 0.823 in our influenza test set; its high performance was significantly boosted by a knowledge-driven heuristic which filtered out locations that are inconsistent with existing locations in the GenBank record. For instance, if a GenBank record contained “Arizona, USA”, and the textual patterns applied by the Text Linker found “Italy”, “Wisconsin” and “Tempe” to be possibly linked to the record, the application of the coherency-checking heuristic would filter out “Italy” and “Wisconsin”, leaving behind only “Tempe”, as the others are inconsistent with the record location of “Arizona, USA”. Since over 99.9% of the influenza records analyzed in our study contained some form of geographic metadata, the use of this heuristic played a major role in improving the performance of the Text Linker. In cases where the geographic metadata in GenBank is absent, the Text Linker tends to struggle considerably more, and the use of more sophisticated methods is needed to help it achieve higher accuracy. Recent works have demonstrated the potential of distance supervision and deep learning methods for enhancing toponym detection in full-text articles related to GenBank records [61]. Similar methods could also be applied for text linkage.

One of the most significant breakthroughs in NLP technology in recent years has been the creation of word embeddings [62] which could be used to represent individual words within machine learning models. Word embeddings are vector representations of words and phrases which can capture their semantic properties, allowing for better language modeling. In order to create word embeddings specifically for our research domain, we developed a Word2Vec model by

downloading and parsing all PMC Open Access articles linked to GenBank. We then utilized the resulting word embeddings to build a convolutional neural network (CNN) incorporating different knowledge-based features. We trained the CNN model using a learning rate of 0.001 on training examples that we generated through distant supervision, instead of manual annotation. The positive training examples were generated based on existing GenBank metadata. For example, if a GenBank record contained the geographic metadata “Tempe, AZ, USA”, all sentences mentioning "Tempe" and "Arizona" in articles related to the record were assumed to be positive. *USA* and other country level-locations were not included to minimize noise. The negative training examples were generated by assuming that all sentences mentioning locations inconsistent with existing GenBank metadata must be negative. For instance, if a GenBank record contained the geographic metadata, "AZ, USA", then sentences mentioning "Alaska" were considered negative while those mentioning "Tempe" were considered neither positive, nor negative. We performed training using all ~50,000 positive examples, and a randomly down-sampled set of ~100,000 negative examples. When tested on a manually annotated test set of 916 GenBank records linked to 21 PMC articles, the trained CNN model achieved a precision, recall, and f-score of 0.51, 0.85 and 0.64 respectively. The rule-based module currently utilized in GeoBoost achieved a precision, recall, and f-score of 0.48, 0.54, and 0.50 respectively on the same set. Please note that these results were obtained without the use of the knowledge-based constraints which limited the output to locations that were consistent with existing GenBank metadata, in contrast to previous evaluations of the Text Linker in GeoBoost. In addition, annotation was also performed in the test set without taking into account existing geographic metadata in GenBank. This allowed us to assess the performance of the two different methods specifically for the task of extracting the LOIH of viruses from full-text publications, without possessing any prior knowledge about the LOIH from GenBank.

Although the CNN model demonstrated promising results on the test set, especially when compared to the rule-based method currently being applied, more rigorous evaluation is needed prior to integrating it within the GeoBoost software. When evaluating the Test Linker, it is important to have an adequately large sample of linked articles in addition to having a large sample of

GenBank records. While the test set contained a fairly large set of 916 GenBank records, the included records were linked to 21 articles only, and therefore it did not provide sufficient insight into the performance of the model. Also, neural network models need to be fine-tuned for optimal performance by adjusting parameters such as the learning rate. Therefore, it is possible that better results, more closely resembling the high performance achieved by state-of-the-art deep learning models for event extraction, could be achieved through different parameter combinations. Moreover, the automatically generated training set contained a fair amount of noise and additional techniques for noise reduction could also help boost performance.

To summarize, future work might include expanding upon the preliminary work done to-date for enhancing the Record Location Extractor and Text Linker and integrating the updated modules within the GeoBoost framework. Also, additional experiments may be performed to evaluate whether existing deep learning models for toponym detection in this domain could add significant value to the GeoBoost system. Moreover, the GeoBoost framework could be adjusted to make it more easily adaptable to GenBank records related to species other than viruses.

6 CONCLUSION

Over the past few decades, we have seen a proliferation of large-scale databases and knowledge-bases, both within and outside the biomedical domain, which provide novel directions for addressing challenging biomedical information extraction tasks. This work presents innovative methods of exploiting the knowledge derived from different resources for geographic information extraction within the biomedical domain, specifically for geo-referencing GenBank records of viruses. Different experiments performed for evaluating these methods consistently demonstrate them to be highly effective for this task.

We used the knowledge-driven geographic IE methods described in this dissertation to develop a publicly available SQL database containing over two million geo-referenced virus GenBank records which could help significantly accelerate public health studies which require the LOIH of virus sequences. In addition, we used these methods to build a command-line program called GeoBoost for extracting, integrating and normalizing the LOIH of viruses based on information present in GenBank and related full-text articles. The GeoBoost framework has been utilized in several research studies by phylogeography researchers and could prove to be a useful tool within the public health domain.

Although developed specifically for processing virus GenBank records, GeoBoost may be easily extended to work with GenBank records related to other species as well and, therefore, has the potential to benefit a wide range of biomedical studies. Moreover, the success of GeoBoost demonstrate the ability of simple knowledge-driven heuristics for addressing complex biomedical problems and underscores the importance of harnessing knowledge from different information sources for biomedical IE. However, GeoBoost, like all other automated systems, is not a perfect system and the integration of state-of-the-art machine learning methods, along with incorporation of additional knowledge-bases, could help further enhance its performance.

REFERENCES

[1] Y. Wang *et al.*, “Clinical information extraction applications: A literature review,” *Journal of Biomedical Informatics*. 2018.

[2] K. Kreimeyer *et al.*, “Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review.,” *J. Biomed. Inform.*, vol. 73, pp. 14–29, 2017.

[3] Y. Garten, A. Coulet, and R. B. Altman, “Recent progress in automatically extracting information from the pharmacogenomic literature.,” *Pharmacogenomics*, vol. 11, pp. 1467–1489, 2010.

[4] M. Simmons, A. Singhal, and Z. Lu, “Text mining for precision medicine: Bringing structure to ehrs and biomedical literature to understand genes and health,” in *Advances in Experimental Medicine and Biology*, 2016.

[5] D. Guin *et al.*, “Global text mining and development of pharmacogenomic knowledge resource for precision medicine,” *Front. Pharmacol.*, vol. 10, p. 839, 2019.

[6] B. Percha, “Biomedical Text Mining From Context,” *Stanford Univ. Diss.*, 2016.

[7] W. Xing *et al.*, “A gene-phenotype relationship extraction pipeline from the biomedical literature using a representation learning approach,” in *Bioinformatics*, 2018.

[8] A. Nikfarjam, A. Sarker, K. O’Connor, R. Ginn, and G. Gonzalez, “Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features,” *J. Am. Med. Informatics Assoc.*, 2015.

[9] A. Sarker *et al.*, “Data and systems for medication-related text classification and concept normalization from Twitter: Insights from the Social Media Mining for Health (SMM4H)-2017 shared task,” *J. Am. Med. Informatics Assoc.*, 2018.

[10] L. L. Knowles and W. P. Maddison, “Statistical phylogeography,” *Mol. Ecol.*, 2002.

[11] R. S. Ostfeld, G. E. Glass, and F. Keesing, “Spatial epidemiology: An emerging (or re-emerging) discipline,” *Trends in Ecology and Evolution*. 2005.

[12] J. L. Gardy and N. J. Loman, “Towards a genomics-informed, real-time, global pathogen surveillance system,” *Nature Reviews Genetics*. 2018.

[13] D. A. Benson *et al.*, “GenBank.,” *Nucleic Acids Res.*, vol. 41, pp. D36-42, Jan. 2013.

[14] M. Scotch *et al.*, “Enhancing phylogeography by improving geographical information from GenBank.,” *J. Biomed. Inform.*, vol. 44 Suppl 1, pp. S44-7, Dec. 2011.

[15] E. C. Holmes, “The phylogeography of human viruses,” *Molecular Ecology*, vol. 13. pp. 745–756, 2004.

[16] N. R. Faria, M. A. Suchard, A. Rambaut, and P. Lemey, “Toward a quantitative understanding of viral phylogeography.,” *Curr. Opin. Virol.*, vol. 1, no. 5, pp. 423–9, Nov. 2011.

[17] M. Woolhouse, F. Scott, Z. Hudson, R. Howey, and M. Chase-Topping, “Human viruses: Discovery and emeraence,” *Philos. Trans. R. Soc. B Biol. Sci.*, 2012.

[18] C. R. Howard and N. F. Fletcher, “Emerging virus diseases: Can we ever expect the unexpected?,” *Emerging Microbes and Infections*. 2012.

[19] P. Lemey, A. Rambaut, A. J. Drummond, and M. A. Suchard, “Bayesian phylogeography finds its roots,” *PLoS Comput. Biol.*, vol. 5, no. 9, 2009.

[20] P. Lemey, A. Rambaut, J. J. Welch, and M. A. Suchard, “Phylogeography takes a relaxed random walk in continuous space and time,” *Mol. Biol. Evol.*, 2010.

[21] K. Brunker *et al.*, “Landscape attributes governing local transmission of an endemic zoonosis: Rabies virus in domestic dogs,” *Mol. Ecol.*, 2018.

[22] J. Raghwani *et al.*, “Endemic dengue associated with the co-circulation of multiple viral lineages and localized density-dependent transmission.,” *PLoS Pathog.*, vol. 7, no. 6, p. e1002064, Jun. 2011.

[23] D. Magee, J. E. Taylor, and M. Scotch, “The Effects of Sampling Location and Predictor Point Estimate Certainty on Posterior Support in Bayesian Phylogeographic Generalized Linear Models,” *Sci. Rep.*, 2018.

[24] D. A. Benson *et al.*, “GenBank,” *Nucleic Acids Res.*, vol. 41, pp. D36-42, 2013.

[25] “Influenza A virus (A/Sendai-H/F005/2006(H3N2)) M1, M2 genes for matrix protein 1, matrix protein 2, partial cds, laboratory strain 3,” Jul-2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/nuccore/AB520868.1. [Accessed: 28-Jul-2019].

[26] “Influenza A virus (A/New York/45/2003(H3N2)) segment 2, complete sequence,” May-2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/nuccore/CY000071.1. [Accessed: 28-Jul-2019].

[27] “Rabies virus G gene for glycoprotein, complete cds, strain: BRdg640,” Jan-2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/nuccore/AB518493.1. [Accessed: 28-Jul-2019].

[28] “Influenza A virus (A/swine/Cambridge/1/1935(H1N1)) nucleoprotein (NP) gene, complete cds,” May-2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/nuccore/M63769.1. [Accessed: 28-Jul-2019].

[29] “Influenza A virus (A/swine/29/1937(H1N1)) nucleoprotein (NP) gene, complete cds,” May-2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/nuccore/M63757.1. [Accessed: 28-Jul-2019].

[30] “Influenza A virus (A/Tianjinheping/SWL313/2009(H1N1)) segment 8 nuclear export protein (NEP) gene, partial cds; and nonstructural protein 1 (NS1) gene, complete cds,” Oct. 2013.

[31] “GeoNames.” [Online]. Available: http://www.geonames.org/. [Accessed: 05-Sep-2013].

[32] “GeoNames Fulltextsearch : bristol.” .

[33] “Eastern equine encephalitis virus strain EEEV/Culiseta melanura/USA/SL13-0764-C/2013, complete genome,” May-2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/nuccore/KX029319.1. [Accessed: 28-Jul-2019].

[34] M. Scotch *et al.*, “Incorporating sampling uncertainty in the geospatial assignment of taxa for virus phylogeography,” *Virus Evol.*, vol. 5, no. 1, p. 43, 2019.

[35] R. J. Fisk, “RSV Genetic Diversity to Global Transmission Dynamics,” The University of

Texas, 2019.

[36] E. S. Chen and I. N. Sarkar, “Towards Structuring Unstructured GenBank Metadata for Enhancing Comparative Biological Studies.,” *AMIA Jt. Summits Transl. Sci. Proc. AMIA Summit Transl. Sci.*, vol. 2011, pp. 6–10, Jan. 2011.

[37] I. N. Sarkar, “Leveraging biomedical ontologies and annotation services to organize microbiome data from Mammalian hosts.,” *AMIA Annu. Symp. Proc.*, vol. 2010, pp. 717–721, 2010.

[38] T. Barrett *et al.*, “BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata.,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D57-63, Jan. 2012.

[39] N. A. O’Leary *et al.*, “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D733-45, Jan. 2016.

[40] A. R. Wattam *et al.*, “Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center,” *Nucleic Acids Res.*, 2017.

[41] E. L. Hatcher *et al.*, “Virus Variation Resource - improved response to emergent viral outbreaks.,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D482–D490, Jan. 2017.

[42] B. E. Pickett *et al.*, “ViPR: An open bioinformatics database and analysis resource for virology research,” *Nucleic Acids Res.*, 2012.

[43] A. Z. Ijaz, T. C. Jeffries, U. Z. Ijaz, K. Hamonts, and B. K. Singh, “Extending SEQenv: a taxa-centric approach to environmental annotations of 16S rDNA sequences,” *PeerJ*, 2017.

[44] R. Bossy, W. Golik, Z. Ratkovic, P. Bessieres, and C. Nedellec, “BioNLP shared Task 2013 – An Overview of the Bacteria Biotope Task,” in *Proceedings of the BioNLP Shared Task Workshop, ACL*, 2013, pp. 161–169.

[45] R. Bossy, J. Jourde, P. Bessières, M. van de Guchte, and C. Nédellec, “BioNLP Shared Task 2011 - Bacteria Biotope,” in *Proceedings of BioNLP Shared Task 2011 Workshop*, 2011, pp. 56–64.

[46] L. Delèger *et al.*, “Overview of the Bacteria Biotope Task at BioNLP Shared Task 2016,” 2016.

[47] D. Weissenbacher, A. Magge, K. O’Connor, M. Scotch, and G. Gonzalez-Hernandez, “SemEval-2019 Task 12: Toponym Resolution in Scientific Papers,” in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 907–916.

[48] X. Wang *et al.*, “DM NLP at SemEval-2019 Task 12: A Pipeline System for Toponym Resolution.”

[49] A. Magge, D. Weissenbacher, A. Sarker, M. Scotch, and G. Gonzalez-Hernandez, “Bi-directional Recurrent Neural Network Models for Geographic Location Extraction in Biomedical Literature,” 2018.

[50] D. Weissenbacher *et al.*, “Knowledge-driven geospatial location resolution for phylogeographic models of virus migration.,” *Bioinformatics*, vol. 31, no. 12, pp. i348–i356, Jun. 2015.

[51] J. Tamames and V. de Lorenzo, “EnvMine: a text-mining system for the automatic extraction of contextual information.,” *BMC Bioinformatics*, vol. 11, p. 294, Jan. 2010.

[52] J. L. Leidner, “Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding,” *SIGIR Forum*, vol. 41, no. 2, pp. 124–126, Dec. 2007.

[53] W. Shen, J. Wang, and J. Han, “Entity linking with a knowledge base: Issues, techniques, and solutions,” *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 443–460, 2015.

[54] D. Buscaldi, “Approaches to Disambiguating Toponyms.”

[55] T. Tahsin *et al.*, “Natural language processing methods for enhancing geographic metadata for phylogeography of zoonotic viruses.,” *AMIA Jt. Summits Transl. Sci. Proc. AMIA Summit Transl. Sci.*, vol. 2014, pp. 102–11, Jan. 2014.

[56] D. Weissenbacher, A. Sarker, T. Tahsin, M. Scotch, and G. Gonzalez, “Extracting geographic locations from the literature for virus phylogeography using supervised and distant supervision methods.,” *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.*, vol. 2017, pp. 114–122, 2017.

[57] D. Weissenbacher *et al.*, “Knowledge-driven geospatial location resolution for phylogeographic models of virus migration,” in *Bioinformatics*, 2015, vol. 31, no. 12, pp. i348–i356.

[58] “Rabies virus G gene for glycoprotein, complete cds, isolate: BKK-TH-171,” Sep. 2016.

[59] “Influenza A virus (A/India/GWL01/2011(H1N1)) segment 2 polymerase PB1 (PB1) gene, complete cds,” Mar. 2013.

[60] S. Sharma *et al.*, “Molecular Epidemiology and Complete Genome Characterization of H1N1pdm Virus from India,” *PLoS One*, 2013.

[61] A. Magge, D. Weissenbacher, A. Sarker, M. Scotch, and G. Gonzalez-Hernandez, “Deep neural networks and distant supervision for geographic location mention extraction,” in *Bioinformatics*, 2018.

[62] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations ofwords and phrases and their compositionality,” in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.