Bioinformatics, 36(20), 2020, 5120–5121
doi: 10.1093/bioinformatics/btaa647
Advance Access Publication Date: 19 July 2020
Applications Note

Data and text mining

GeoBoost2: a natural languageprocessing pipeline for GenBank metadata enrichment for virus phylogeography

Arjun Magge[1,2,3], Davy Weissenbacher[3], Karen O'Connor[3], Tasnia Tahsin[1], Graciela Gonzalez-Hernandez[3] and Matthew Scotch[1,2,*]

[1]College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA, [2]Biodesign Center for Environmental Health Engineering, Biodesign Institute, Arizona State University, Tempe, AZ 85287, USA and [3]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine,University of Pennsylvania, Philadelphia, PA 19104, USA

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on March 27, 2020; revised on July 3, 2020; editorial decision on July 12, 2020; accepted on July 13, 2020

Abstract

Summary: We present GeoBoost2, a natural language-processing pipeline for extracting the location of infected hosts for enriching metadata in nucleotide sequences repositories like National Center of Biotechnology Information's GenBank for downstream analysis including phylogeography and genomic epidemiology. The increasing number of pathogen sequences requires complementary information extraction methods for focused research, including surveillance within countries and between borders. In this article, we describe the enhancements from our earlier release including improvement in end-to-end extraction performance and speed, availability of a fully functional web-interface and state-of-the-art methods for location extraction using deep learning.

Availability and implementation: Application is freely available on the web at https://zodo.asu.edu/geoboost2. Source code, usage examples and annotated data for GeoBoost2 is freely available at https://github.com/ZooPhy/geoboost2.

Contact: matthew.scotch@asu.edu

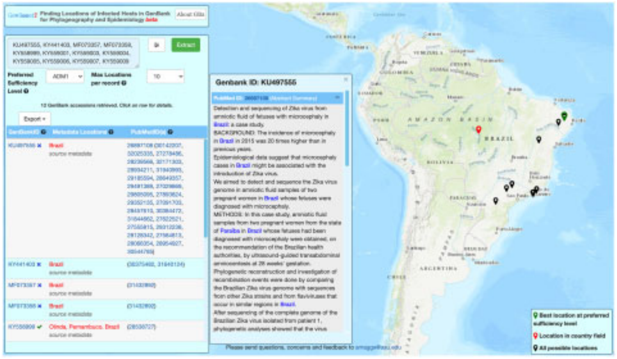
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Molecular sequences play a vital role in conducting phylogenetic, phylogeographic and epidemiological studies to understand the dynamic nature of evolution and migration of pathogens across countries and continents. The National Center of Biotechnology Information (NCBI) maintains GenBank (Benson *et al.*, 2018), which is one of the largest comprehensive databases of nucleotide sequences available to the public. As of July 2020, GenBank contains 217 million entries (NCBI, 2020a) with over 3 million viral sequences reported in the latest release notes (NCBI, 2020b). The availability of such a database supports research in various domains of public health, particularly infectious diseases such as Ebola, Zika and most recently SARS-CoV-2 (Dudas *et al.*, 2017; Lai *et al.*, 2020; Pybus *et al.*, 2012). However, the quality of geographic metadata about the location of infected hosts (LOIH) that is readily available at the individual record level may be insufficient for studies conducted at the state/province levels within the country (Scotch *et al.*, 2011; Tahsin *et al.*, 2014). The presence of detailed geographic metadata is crucial not just for epidemiological studies, but also in retrospective genomic studies by the wider scientific community.

Geographic metadata about the infected host is not required when submitting a sequence to GenBank. The database offers a *Features* table which includes both mandatory and optional qualifiers (Benson *et al.*, 2018; INSDC, 2019). Geographic metadata is amongst the optional qualifiers including *lat_lon* for the approximate coordinates, and *country* for named locations. Among the over 3 million viral sequences available (NCBI, 2020b), only about 1% of the records contained the infected host's coordinates in the lat_lon field and only 26% contained host information more specific than a country in the *country* field. Such unavailability of detailed metadata in GenBank creates barriers for phylogeographic and genomic epidemiology at a local level. Researchers are then required to manually analyze other metadata fields in the record and/or review any associated PubMed articles. If no additional metadata is found, then the researcher might decide to exclude these records from the study altogether, reducing the sample size of the study and potentially introducing bias.

GeoBoost2 provides a framework to automate this manual extraction process where the individual metadata fields are analyzed with the objective of extracting the LOIH from associated records. GeoBoost2 improves over its predecessor GeoBoost (Tahsin *et al.*,


GeoBoost2 Finding Locations of Infected Hosts in GenBank for Phylogeography and Epidemiology beta

KU497555, KY441403, MF073357, MF073358, KY558989, KY558991, KY558993, KY558994, KY558995, KY559006, KY559007, KY559009

Preferred Sufficiency Level: ADM1

Max Locations per record: 10

12 GenBank accessions retrieved. Click on row for details.

GenBankID	Metadata Location	PubMedID(s)
KU497555 ✗	Brazil source metadata	26887108 (30142207, 32025335, 27279496, 28239568, 32171303, 28004211, 31940993, 29185594, 28649357, 29491388, 27029866, 29605095, 27893624, 29352135, 27091703, 28487613, 30384472, 31644962, 27522521, 27558815, 29312236, 28128342, 27564813, 28086354, 28954927, 30544765)
KY441403 ✗	Brazil source metadata	(30375462, 31640124)
MF073357 ✗	Brazil source metadata	(31432800)
MF073358 ✗	Brazil source metadata	(31432800)
KY558989 ✓	Olinda, Pernambuco, Brazil source metadata	(28536727)

Genbank ID: KU497555

PubMed ID: 26897108 (Abstract Summary)

Detection and sequencing of Zika virus from amniotic fluid of fetuses with microcephaly in Brazil: a case study.
BACKGROUND: The incidence of microcephaly in Brazil in 2015 was 20 times higher than in previous years.
Epidemiological data suggest that microcephaly cases in Brazil might be associated with the introduction of Zika virus.
We aimed to detect and sequence the Zika virus genome in amniotic fluid samples of two pregnant women in Brazil whose fetuses were diagnosed with microcephaly.
METHODS: In this case study, amniotic fluid samples from two pregnant women from the state of Paraiba in Brazil whose fetuses had been diagnosed with microcephaly were obtained, on the recommendation of the Brazilian health authorities, by ultrasound-guided transabdominal amniocentesis at 28 weeks' gestation.
Phylogenetic reconstruction and investigation of recombination events were done by comparing the Brazilian Zika virus genome with sequences from other Zika strains and from flaviviruses that occur in similar regions in Brazil.
After sequencing of the complete genome of the Brazilian Zika virus isolated from patient 1, phylogenetic analyses showed that the virus

Best location at preferred sufficiency level

Location in country field

All possible locations

Please send questions, concerns and feedback to amagge@asu.edu

Fig. 1. Screenshot from the GeoBoost2 website. In this example, the user enters GenBank accession IDs for Zika virus and designates sufficiency level in terms of administrative divisions (GeoNames, 2020a), such as ADM1 for states/provinces, ADM2 for county and maximum number of possible locations to be displayed per record for the search. Upon submission of the request, GeoBoost2 extracts locations from GenBank record metadata. For each record where the sufficiency level is not met, GeoBoost2 checks associated PubMed abstracts/open access articles. The system then displays all possible locations it extracted with details available on hover over the pins on the map. The user can then export the data in csv, tsv or json formats. For record KU497555 (Calvet *et al.*, 2016), only the country information was available in the metadata but a finer location; in this case, the state of Paraiba was found in one of the linked papers

2016, 2018) in extraction performance by over 35% when evaluated on two corpora using advanced data mining methods on the linked PubMed articles to enrich the geospatial metadata. Overall, GeoBoost2 achieved 90% accuracy in resolving the LOIH in GenBank metadata and 57% accuracy in resolving LOIH extraction from associated PubMed articles. To the best of our knowledge, GeoBoost and GeoBoost2 are the only systems that using natural language-processing (NLP) techniques to extract LOIH from articles cited in GenBank accessions. In Supplementary Information, we describe in detail our methods and evaluation of GeoBoost2. We also provide a screenshot of the current version of the interface (Fig. 1).

GeoBoost2 includes:

1. A state-of-the-art deep-learning NLP algorithm trained on manually annotated geographic location mentions in PubMed Central Open Access articles (Magge *et al.*, 2018, 2019). All geographic location mentions are disambiguated and resolved to a unique identifier in GeoNames (2020a,b), a database containing 12 million locations across the globe.
2. A Python 3.7 framework implementation (replacing a Java-based framework) for continuous improvement with deep-learning and machine-learning methods for information extraction.
3. A Web-based interface with a map view that accepts as input any GenBank accessions (not limited to viruses) and provides features to export results. In addition to accepting GenBank accession IDs, the tool can also accept PubMed IDs or raw text captured from an article for mining geographic locations.
4. An application programming interface (API) for use of the results in downstream applications. In addition to mining PubMed articles directly linked in the GenBank accessions, GeoBoost2 also mines geographic locations from additional PubMed articles and their respective Supplementary Information that have cited the GenBank accessions in their studies. All data retrieval functionalities in the tool rely on APIs provided by NCBI, ensuring the latest available information.

Results from GeoBoost2 can be used for Bayesian discrete phylogeography on ZooPhy (Scotch *et al.*, 2010, 2019b; ZooPhy, 2020). Here, the probabilities for potential LOIH generated by GeoBoost2 can be used as sampling uncertainties (Scotch *et al.*, 2019a) for the taxa in phylogeographic studies implemented using BEAST (Suchard *et al.*, 2018).

We plan to extend our information extraction and normalization efforts to additional optional qualifiers such as *collection_date*, *host* and *isolation_source*. We also plan to validate the performance of the tool on other pathogens such as bacteria and parasites.

With the growing concern over emerging and re-emerging pathogens, a publicly available, free tool like GeoBoost2 will facilitate public health surveillance and genomic epidemiology.

Acknowledgements

The authors thank Megan Rorison, Briana Trevino, Robert Rivera, Rachel Beard and Mari Firago for annotating the corpus. The authors acknowledge the NVIDIA Corporation for the donation of the Titan Xp GPU used for training deep-learning models in this research.

Funding

The research presented was supported by the National Institutes of Health (NIH) through NIAID Award [R01AI117011 to G.G.-H. and M.S.] and NLM Award [R01LM012080 to M.S.]. The content is solely the responsibility of the authors and does not necessarily represent the views of the NIH.

Conflict of Interest: none declared.

References

Benson,D. *et al.* (2018) GenBank. *Nucleic Acids Res.*, **46**, D41–D47.

Calvet,G. *et al.* (2016) Zika virus isolate brazil-zkv2015, complete genome. GenBank database, https://www.ncbi.nlm.nih.gov/nuccore/ku497555, (26 July 2020, date last accessed).

Dudas,G. *et al.* (2017) Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*, **544**, 309–315.

GeoNames. (2020a) Feature codes. https://www.geonames.org/export/codes.html, (26 July 2020, date last accessed).

GeoNames. (2020b) https://www.geonames.org/about.html, (26 July 2020, date last accessed).

INSDC (2019) http://www.insdc.org/files/feature_table.html#2.2.

Lai,C. *et al.* (2020) Severe acute respiratory syndrome coronavirus 2 (sars-cov-2) and coronavirus disease-2019 (covid-19): the epidemic and the challenges. *Int. J. Antimicrob. Agents*, **55**, 105924.

Magge,A. *et al.* (2018) Deep neural networks and distant supervision for geographic location mention extraction. *Bioinformatics*, **34**, i565–i573.

Magge,A. *et al.* (2019) Bi-directional recurrent neural network models for geographic location extraction in biomedical literature. *Pac. Symp. Biocomput.*, **24**, 100–111.

NCBI (2020a) Genbank statistics. https://www.ncbi.nlm.nih.gov/genbank/statistics/, (26 July 2020, date last accessed).

NCBI (2020b) Release notes for genbank release 237. https://www.ncbi.nlm.nih.gov/genbank/release/237/, (26 July 2020, date last accessed).

Pybus,O.G. *et al.* (2012) Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl. Acad. Sci. USA*, **109**, 15066–15071.

Scotch,M. *et al.* (2010) At the intersection of public-health informatics and bioinformatics: using advanced web technologies for phylogeography. *Epidemiology (Cambridge, Mass.)*, **21**, 764–768.

Scotch,M. *et al.* (2011) Enhancing phylogeography by improving geographical information from genbank. *J. Biomed. Inform.*, **44**, S44–7.

Scotch,M. *et al.* (2019a) Incorporating sampling uncertainty in the geospatial assignment of taxa for virus phylogeography. *Virus Evol.*, **5**, vey043.

Scotch,M. *et al.* (2019b) Zoophy: a bioinformatics pipeline for virus phylogeography and surveillance. *Online J. Public Health Inf.*, **11**, e301.

Suchard,M.A. *et al.* (2018) Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus Evol.*, **4**, vey016.

Tahsin,T. *et al.* (2014) Natural language processing methods for enhancing geographic metadata for phylogeography of zoonotic viruses.AMIA Jt Summits Transl Sci Proc., **2014**, 102–111.

Tahsin,T. *et al.* (2016) A high-precision rule-based extraction system for expanding geospatial metadata in genbank records. *J. Am. Med. Inform. Assoc.*, **23**, 934–941.

Tahsin,T. *et al.* (2018) Geoboost: accelerating research involving the geospatial metadata of virus genbank records. *Bioinformatics*, **34**, 1606–1608.

ZooPhy. (2020) https://zodo.asu.edu/zoophy/.