NIH Public Access
Author Manuscript
Epidemiology. Author manuscript; available in PMC 2011 November 1.

Published in final edited form as:
Epidemiology. 2010 November ; 21(6): 764–768. doi:10.1097/EDE.0b013e3181f534dd.

At the intersection of public-health informatics and bioinformatics: Using Advanced Web Technologies for Phylogeography

Matthew Scotch[1], **Changjiang Mei**[1], **Cynthia Brandt**[1], **Indra Neil Sarkar**[2,3], and **Kei Cheung**[1]
[1]Yale Center for Medical Informatics, Yale University, New Haven, CT USA
[2]Center for Clinical and Translational Science, University of Vermont, Burlington, VT USA
[3]Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, VT USA

Web technologies that promote information-sharing and collaboration are known as Web 2.0.[1]. These technologies include the popular social networks and wikis, Web services (Web application programming interfaces [APIs] such as SOAP or REST) that support interoperability between computers,[2] and mashup and workflow environments such as Yahoo! Pipes (Yahoo! Inc., Sunnyvale, CA USA) and Taverna (University of Manchester, Manchester, UK). Geospatial applications such as Google Earth (Google, Mountain View, CA USA) have facilitated visualization and analysis of geo-referenced data over the Web. Another advanced Web environment is the Semantic Web, which uses the Resource Description Framework (RDF) and the Ontology Web Language (OWL). These frameworks enable Web resources to be annotated and defined in a way that facilitates information-sharing and knowledge discovery.[3,4]

Phylogeography and Translational Public Health

Health sciences can benefit from the use of advanced Web technologies. One example is phylogeography, a sub-discipline of biogeography that focuses on the geographical spread of genealogical lineages of species ranging from viruses to mammels.[5] In this field, both spatial and temporal factors are considered when studying the evolutionary lineage of a species. Phylogenetic trees (hierarchical models of evolutionary history) are often considered within a spatial context such as a map. Thus phylogenetics and molecular evolution are cornerstones of this field.

Because of their often-shorter genomes and rapid transformations, RNA viruses in particular lend themselves to phylogeography.[6] For example, Biek et al.[7] studied the distribution of rabies virus in raccoons in the northeast United States. The authors examined 30 years of data in order to model the viral population based on an initial wave of infection.[7] A maximum clade credibility tree was constructed from Bayesian coalescent analysis and integrated with information on location. Seven distinct genetic lineages were found. The integration of this evolutionary tree with the map of northeastern United States enabled the researchers to examine the geographic spread of the various strains over time. Use of Bayesian skyline plots then enabled researchers to estimate the number of rabies infections over time.[7]

Address for correspondence: Matthew Scotch, Department of Biomedical Informatics, Arizona State University, 425 N. 5th St., Phoenix, AZ 85004, Phone: 602.827.2500, matthew.scotch@yale.edu.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Phylogeography of zoonotic viruses (animal-human transmission) can support public health and disease surveillance. For example, it can help epidemiologists describe which animal hosts most affect virus propagation, in a particular geographical area and time-frame what migration path—a particular virus has followed, where the virus is going, and the estimated viral population within a specific animal host. In these ways data originally generated for genetic research can help inform public health decision-making (Figure 1).

Large amounts of genetic and biodiversity data are being collected and curated in electronic databases.[8] A combination of informatics techniques, such as data mining, advanced Web technology (including Web 2.0 and the Semantic Web), as well as techniques for studying molecular evolution, such as phylogenetic analysis, can be used to integrate and process the data for phylogeography. Epidemiologists can then use this information to answer public health questions. As data on genetics, biodiversity, and geography are is increasingly introduced into public health practice through the use of advanced Web technology, epidemiologists will be able to identify the most susceptible at-risk animals and target them in the surveillance of particular zoonotic diseases.

Collaboration among phylogeography, bioinformatics and public health informatics can lead to new integrated knowledge across traditionally-disjointed disciplines. Many researchers of phylogeography have expressed a need for the field to take a more analytic framework, rather than simply describe the geographic dispersal of species.[10] Web 2.0 and Semantic Web applications offer potential solutions to this problem. Web 2.0 contains mashup tools that can combine and process disparate Web services and data streams, reducing the need for new programming.

ZooPhy

Advanced Web technologies have gained attention for being able to integrate diverse public health and genetic data. At Yale University, we are developing a framework for the phylogeography of infectious zoonotic agents that represents an intersection of bioinformatics and public health informatics. The purpose is to combine genetic sequence data, develop evolutionary models (phylogenies) of the data, and combine the results with traditional reportable-disease data collected by health departments. This mashup of data could enable epidemiologists to better understand disease migration in animal hosts, viral population in these hosts, and the impact of migration and viral population on human risk of infection and disease. We describe here the preliminary design and development of our system, ZooPhy, using advanced Web technology.

Figure 2 shows the back-end architecture of ZooPhy that contains advanced Web technologies. The left side of the figure shows the retrieval of various data sources. For geographic data, Geonames.org was utilized. This Web site contains many Web services that return geographic locations, including a Hierarchy Web service that identifies the latitude and longitude of a location, as well as its “parent” locations.[11] For example, a query for Bedford, NH might return the latitude and longitude for this town, as well as its county (Hillsborough), its state (New Hampshire), county (United States), etc. For species data, taxonomic information was retrieved from the National Center for Biotechnology Information (NCBI). NCBI has Web services (“E-Utilities”) that connect to its collection of “Entrez” databases that include the NCBI Taxonomy database,[12] which contains a hierarchy of organisms. In addition to this database, the E-Utilities include the GenBank database, which contains INSDC (International Nucleotide Sequence Database Consortium) accession numbers for each molecular sequence data entry. The INSDC is an international collaboration led by United States, European Union, and Asian partners that establishes guidelines for overseeing and accessing the largest publicly-available resource for molecular sequence data. Data integration of the taxonomic, genetic, and geographic information was

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

done in SQL Server (Microsoft Corporation, Redmond, WA). Plans include adding additional NCBI resources such as the Genome and Gene database.

ZooPhy is being developed within a work-flow architecture—a system of processes, like an assembly line, that take as input results from the process immediately before it. Yahoo! Pipes and many other Web 2.0 mashup tools can support this environment. For development of our workflow, we are using Taverna software, which has been used in many recent bioinformatics projects.[13–15] ZooPhy takes query terms as input, as well as an e-mail address to send results back to the user. Query terms are searched against the SQL Server database, and the series of workflow processes begins. In order to address the research questions posed earlier, it is essential to develop a valid model of evolution using phylogenetic approaches. As such, ZooPhy's workflow contains processes that perform nucleotide-sequence alignment and development of phylogenetic trees. Currently, ZooPhy utilizes Web services for alignment using the ClustalW[16] Web service, and Bayesian inference phylogeny using a Web service that implements Bayesian Evolutionary Analysis by Sampling Trees (BEAST) software.[17] Additional phylogenetic approaches, such as maximum likelihood and maximum parsimony, will likely be added to ZooPhy in the near future.

Geographic visualization and analysis are an important component of advanced Web technology. These features have gained popularity through products such as Google Earth, with the ability to mashup different geographic layers to produce a map or globe. Once ZooPhy's workflow has completed, the user will be sent an e-mail that contains a link to the results. Through this link, the user will be able to see a geographic view of the evolution of the zoonotic disease within the animal host (Figure 3).

Conclusion

Web 2.0 and the Semantic Web (3.0) provide great opportunities for biomedical informatics research and the development of informatics tools and resources to address problems across the full spectrum of health science research. One example is ZooPhy, an automated workflow for phylogeography. This system may be useful for by epidemiologists who conduct surveillance and analysis of zoonotic (animal-human) agents. In addition to genetic, taxonomic, and geographical data, ZooPhy will include traditional public-health data collected by health departments. Through phylogenetics, data-mining, and machine-learning approaches, this system may help epidemiologists better understand the migration of various zoonotic diseases in animal hosts, estimate of the viral population growth within these hosts, and calculate risk to humans within a defined geographic area.

Acknowledgments

Funding: Supported in part by The National Library of Medicine (NLM) grants 5K99LM009825-02 and ARRA supplement 3K99LM09825-02S1 to Matthew Scotch. Indra Neil Sarkar is supported in part by DHHS/NIH/NLM R01LM009725.

The authors thank William Piel from the Yale Peabody Museum for his advice and knowledge on the evolutionary biology techniques of the system.

References

1. O'Reilly, T. What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. 2005. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html?
2. W3C. Web Services Architecture: W3C Working Group Note. Vol. 2010. February 11. 2004

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

3. Post LJ, Roos M, Marshall MS, van Driel R, Breit TM. A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data. Bioinformatics. 2007; 23(22):3080–7. [PubMed: 17881406]

4. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. 2008; 41(5):706–16. [PubMed: 18472304]

5. Avise, JC. Phylogeography: the history and formation of species. Cambridge, Mass: Harvard University Press; 2000.

6. Holmes EC. The phylogeography of human viruses. Mol Ecol. 2004; 13(4):745–56. [PubMed: 15012753]

7. Biek R, Henderson JC, Waller LA, Rupprecht CE, Real LA. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. Proc Natl Acad Sci U S A. 2007; 104(19):7993–8. [PubMed: 17470818]

8. Sarkar IN. Biodiversity informatics: organizing and linking information across the spectrum of life. Brief Bioinform. 2007; 8(5):347–57. [PubMed: 17704120]

9. Rabinowitz P, Scotch M, Conti L. Human and animal sentinels for shared health risks. Vet Ital. 2009; 45(1):23–4. [PubMed: 20148187]

10. Kidd DM, Ritchie DM. Phylogeographic information systems: putting the geography into phylogeography. Journal of Biogeography. 2006; 33:1851–65.

11. Geonames Hierarchy Webservice. 2009. http://www.geonames.org/export/place-hierarchy.html#hierarchy

12. NCBI. Taxonomy Database. 2009. http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy

13. Lanzen A, Oinn T. The Taverna Interaction Service: enabling manual interaction in workflows. Bioinformatics. 2008; 24(8):1118–20. [PubMed: 18337261]

14. Li P, Oinn T, Soiland S, Kell DB. Automated manipulation of systems biology models using libSBML within Taverna workflows. Bioinformatics. 2008; 24(2):287–9. [PubMed: 18056069]

15. Krabbenhoft HN, Moller S, Bayer D. Integrating ARC grid middleware with Taverna workflows. Bioinformatics. 2008; 24(9):1221–2. [PubMed: 18353787]

16. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994; 22(22):4673–80. [PubMed: 7984417]

17. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007; 7:214. [PubMed: 17996036]

18. Wallace RG, Fitch WM. Influenza A H5N1 immigration is filtered out at some international borders. PLoS One. 2008; 3(2):e1697. [PubMed: 18301773]

Biography

MATTHEW SCOTCH is an Assistant Professor in Biomedical Informatics at Arizona State University with a research focus in public health informatics and bioinformatics. He was the recipient of an NIH Pathway to Independence (K99/R00) career development award through the National Library of Medicine. His research is focused on combining disparate types of data including genetic, taxonomic, geographical, and public health case data for surveillance of zoonotic (animal-human) diseases.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

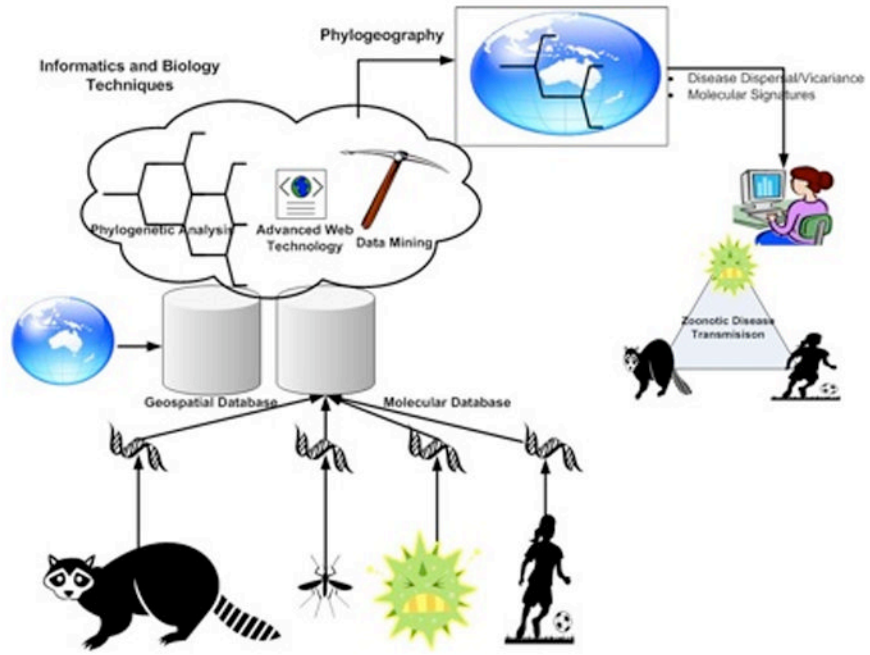

Figure 1.
Phylogeography in support of Translational Public Health. Integration of genetic, geospatial, biodeversity data with informatics methods and techniques for studying molecular evolution can enable epidemiologists to better understand molecular signatures of zoonotic disease dynamics. Adapted from paper by Rabinowitz, Scotch and Conti.[9]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

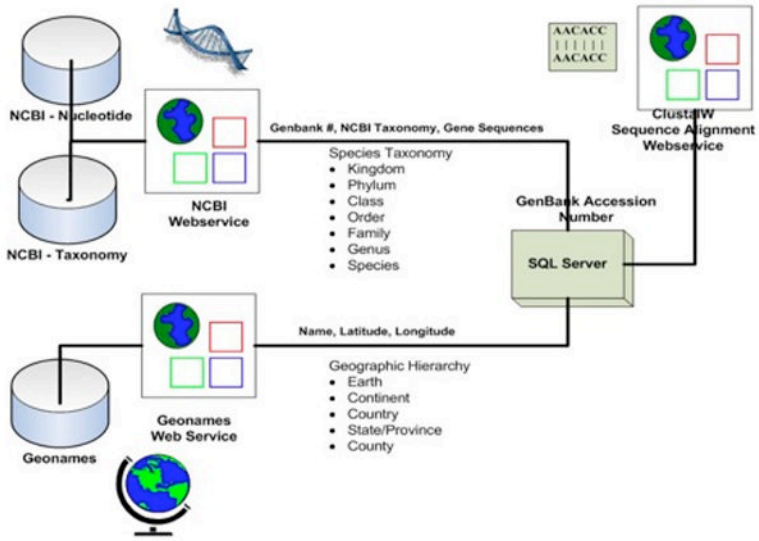

Figure 2.
Back-end Architecture for ZooPhy.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

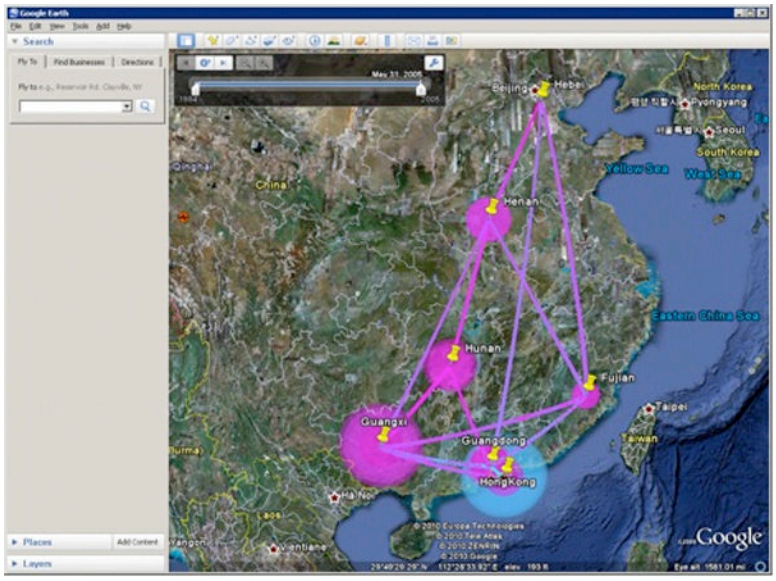

Figure 3.
Output of ZooPhy workflow shown in Google Earth. In this example, the migration of Avian Influenza in China from a variety of hosts is shown. Data from paper by Wallace and Fitch.[18]