MTR 090064
MITRE TECHNICAL REPORT

Genomics for Bioforensics

MITRE Sponsored Research Final Report

Marc Colosimo
Lynette Hirschman
Meredith Keybl
Joanne Luciano
Scott Mardis
Matthew Peterson

December 2008

MTR 090064
MITRE TECHNICAL REPORT

Genomics for Bioforensics

MITRE Sponsored Research Final Report

Marc Colosimo
Lynette Hirschman
Meredith Keybl
Joanne Luciano
Scott Mardis
Matthew Peterson

December 2008

The views, opinions and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.
This document has been publically released, 09-0455.

Abstract

The goal of the Genomics for Bioforensics project (FY06-FY08) has been to explore the application of genomics to the challenge of attribution for biological organisms: given a sample of an agent, can we use genomics to (help) determine whether the new sample is an endemic strain, a strain introduced from another place or time, or a novel (possibly engineered) strain? Specifically, given sequence data from a new sample, we have developed procedures to compare this “unknown” sequence to a reference database of sequences from previously collected samples and their associated metadata. To do this, we have created a reference database for a specific organism (influenza) and developed procedures to create a Microbial Forensics Workbench, which enables the user to compare the “unknown” strain with the strains in the Reference Database. The Workbench provides a novel automated method for clustering the “unknown” strain with the most similar strains in the Reference Database. This method for genotyping uses Complete Composition Vectors and Affinity Propagation Clustering (submitted for publication: Peterson et al., Bioinformatics). The Workbench also supports several visualization techniques, including display of a color-coded phylogenetic tree (using TreeViewJ, Colosimo et al., BMC Bioinformatics, also developed under this project), as well as map and timeline displays based on the Simile open source software.

Table of Contents

List of Figures

List of Tables

1 Introduction

The field of microbial forensics is undergoing a revolution due to advances in sequencing technology. While older methods of distinguishing strains and samples for forensics relied heavily on partial sequencing and searches for diagnostic features such as differing numbers of repeats, the cost and speed of current sequencing technology now makes it cost-effective to sequence entire genomes as the basis for forensics analysis. The current commercial next-generation (second generation) sequencers, such as Roche's (454) GS FLX Genome Analyzer, Illumina's Solexa 1G, and Applied Biosystem's SOLiD, have the ability to rapidly sequence the genome of bacteria. With one Solexa machine, it is possible to find all of the SNPs (single nucleotide polymorphisms) in 16 bacteria in a week. Third generation sequencers will have the potential to sequence whole genomes in one experiment.

This breakthrough in sequencing technology opens new possibilities and creates new challenges for the collection, integration, and analysis of genomics-based microbial forensics data. A major challenge will be to collect and organize genomic data sets that characterize the pathogens of interest that can help to distinguish, at a high level, naturally occurring outbreaks from intentionally distributed pathogens, and at a finer level, to inform the distinction of the probable origin of a particular sample, including the possibility of engineered pathogens.

To use the genome of a pathogen in a forensics context, critical components include:

- Models of microbial evolution, including pathogen life cycle, rate of evolution, global distribution of its strains or genotypes, its internal genomic structure (number of genes and virulence factors, pathogenicity islands, plasmids, phages).
- Capture in an integrated database of both genomic sequence data and metadata for a sufficient number of samples of an organism, distributed over a suitable geographic range and temporal interval, for each pathogen of interest.
- Data representation standards for capturing, exchanging, and searching rich metadata, including temporal (time of collection) and geospatial distribution of samples (place of collection), host, pathogenicity, habitat, method of sample collection and sample processing, and annotations (for genes, virulence factors).
- Algorithms to analyze genome level differences among pathogen strains over short time periods (days to years), for rapidly increasing numbers of samples.
- Visualization of information from thousands of sequenced genomes for use in the analytical process, including support for phylogenetic analysis, genotyping, signature identification, etc.

The Genomics for Bioforensics project has focused on this challenge: how to organize genomic sequence data from multiple strains of an organism, in order to support microbial forensics. We have chosen to focus on influenza because this is a serious threat to the world’s health, and because at the time that we started this project, it was the only organism for which there were publicly available sequence data for hundreds (now thousands) of strains. However, in parallel, we have also explored the application of our methods to other areas. In FY07, we prepared a report on the availability of sequence data for agricultural organisms; in FY08, we applied the genotyping approach developed under this project to microbes with much larger genomes (mycobacterium, streptomyces, actinobacteria).

Section 2 of this report covers the project objects and describes the project deliverables. These deliverables include the Microbial Forensics Workbench as well as software deliverables, resources and algorithms developed under this project. Sections 3-5 describe the three major components of the Workbench:

- The Microbial Forensics Reference Database (MFDB) stores sequence data and metadata for a specific organism.
- A novel sequence analysis and genotyping software to compare strain similarity in “genome space.”
- A visualization prototype that provides tools to enable analysts to visualize a complex space in terms of geospatial and temporal dimensions or phylogenetic relations.

We conclude with a section on Lessons Learned and Impact.

2 Overview

2.1 Objective

Microbial forensics, or Bioforensics, is defined as a scientific discipline dedicated to analyzing evidence from a bioterrorism act, biocrime, or inadvertent microorganism/toxin release for attribution purposes[1]. The scientific underpinnings of this discipline draw heavily on bioinformatics and epidemiological modeling techniques for identification and attribution of these agents. Bioforensics is used to aid in determining the cause of a bio-event – for example, whether or not the outbreak was natural, introduced, or engineered. This problem requires multiple sources of information: both the genomic sequence data and the associated metadata, such as date/location of sampling, host information, and host/patient outcome.

This MSR has focused on developing new methods and tools for microbial forensics based on the rapid advances in genome sequencing. In addition, we have worked with the primary genome sequence data collectors and organizers of sequence data repositories to develop guidelines for data collection in order to identify the metadata necessary for genome-based forensic analysis, so that the metadata can be captured and encoded along with the sequence data.

2.2 Approach

Our approach has focused on the development of an end-to-end system for genome-based bioforensics analysis. This system, called the Microbial Forensics Workbench, has been designed to provide analysts with the tools they need to make attribution decisions based on genomic data and associated metadata. The workbench combines three major activities under this MSR:

- The development of a reference database for storage and retrieval of organism-specific genome sequence data and metadata.
- The development of genome sequence analysis tools for quickly and accurately understanding the relationships between the genetic sequences of samples.
- The development of visualization tools to support analysts in making attribution decisions.

The first step in designing this system was to identify a model pathogen for analysis. Influenza A was chosen as the model organism based on the number of full genome sequences available, as well as the interest in influenza A from both the research community and our sponsors, due to its pandemic potential. We also surveyed the publicly accessible

[1] Budowle, B., Schutzer, S.E., Einseln, A., Kelley, L.C., Walsh, A.C., Smith, J.A., Marrone, B.L., Robertson, J. and Campos, J. (2003) Public health. Building microbial forensics as a response to bioterrorism, *Science*, **301**, 1852-1853.

sequence databases containing data and metadata on plant and animal pathogens to identify possible pathogens for future analysis. This information was written up as a white paper and sent to interested parties in both government and industry.

The main workflow of the system is shown below in Figure 1. Once the organism that caused a bio-event has been identified, and the sequence data and metadata are collected, they are loaded into the reference database. Next, a set of "background" sequences are identified. These can either be chosen based on their genetic similarity to the sequence of interest or by their metadata. At this point, the sequences are genotyped, or grouped on the basis of their sequences. Visualizations are then created, combining the inferred genotypic information with the metadata. Each part of this workflow is discussed in detail in the following sections.

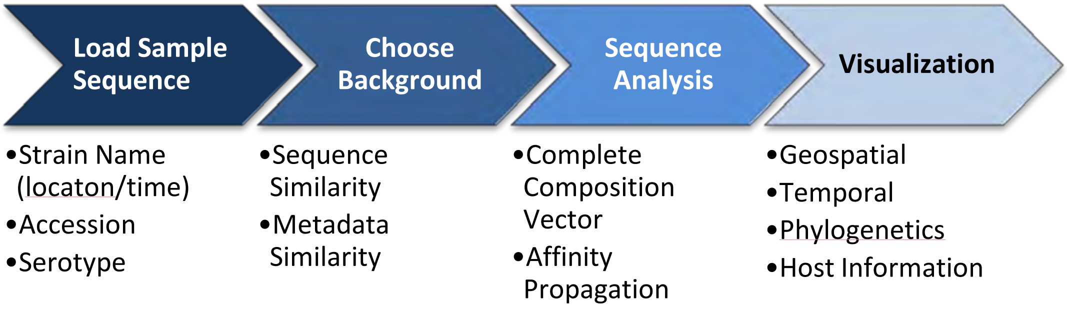


Figure 1: Overview of the Microbial Forensics Workbench

Because such a system is dependent on quality metadata, we have partnered with several groups responsible for the collection and storage of influenza data to produce an ontology for influenza research and surveillance (InfluenzO). This ontology will help define guidelines for the collection of influenza data and metadata for forensic purposes, as well as provide analysts with a model for integrating data from various sources.

2.3 Deliverables

There are three classes of deliverables under this project: software, algorithms and resources.

2.3.1 Software

- TreeViewJ: An open source software for visualization of phylogenetic trees:
    - Described in published article, Matthew Peterson and Marc Colosimo. "TreeViewJ: An application for viewing and analyzing phylogenetic trees." *Source Code for Biology and Medicine*, October 2007.
    - ~500 downloads of software, paper viewed 1,822 times.

2.3.2 Algorithms

- Genotyping software based on Complete Composition Vectors (CCV) and Affinity Propagation for automated clustering of genomic (or proteomic) sequence data:
    - This approach has been demonstrated to scale to hundreds of sequences and to large genomes (millions of base pairs).
    - This work has been described in journal article now under review: Peterson et al., "Automated genotyping via complete composition vectors and affinity propagation clustering." In preparation.
- Extraction of geospatial information from influenza strain names
- Strain name parsing code has been publicly released:
    - This algorithm depends on Integrated Geospatial Database (IGDB), done under government funding and the code has not (yet) approved for release.

2.3.3 Resources

- Influenza ontology (InfluenzO), a collaboration between MITRE, BioHealthBase (U. Texas), and Gemina (U. Maryland).
- Reference database for influenza has been created and documented, but will not be maintained unless additional funding is made available.

3 Developing a Reference Database for Influenza Data and Metadata

3.1 Database Development

3.1.1 Requirements

The task of the Microbial Forensics Reference Database (MFDB) is to be a repository of influenza sequences and metadata supporting attribution analysis. The data and metadata come from several sources. Its primary data is derived from GenBank[2], though that can be augmented with metadata obtained from related influenza databases such as the "Influenza Virus Resource" at the National Center for Biotechnology Information (NCBI). It must also be possible to house private sequence data, not obtained from public sources, which may be part of a current event or private sample collection.

Because public sources of data are regularly updated, the MFDB must support frequent incremental updates. During an update, it will be important to note changes to existing sequences and metadata in order to inform analysts and update models to be current with new information.

One of the greatest challenges in creating and maintaining the database is the management of identifiers that specify strains, sequences, and the source records from which they are derived. Each sequence and strain has several potential identifiers and it is necessary to understand many subtleties of the GenBank identifier schemes to properly manage the import and maintenance of reference data. Appendix A: Sequence Identifiers in GenBank gives a summary of GenBank identifier management.

3.1.2 MFDB Schema

Our schema has a notably different structure than GenBank. We have tables for strains and strain-related information. In GenBank, only strain names in each sequence record link sequences of a common strain together. Therefore, these strain names are also the only indication that the various sequences come from the same source. Taxon identifiers are maintained in GenBank. For influenza, each sample is assigned to its own unique taxon according to influenza naming conventions. These can be used to select sequences from a common sample (e.g., the individual genes) but no significant information is maintained in association with taxon identifiers.

The most significant difference between the BRDB and GenBank is in our treatment of sequences. In the MFDB, the primary sequence table holds only *the coding sequences*

[2] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2008) GenBank, *Nucleic Acids Res*, **36**, D25-30.

(nucleotide sequences that code for proteins), not the raw nucleotide sequences that make up the viral genome.

In GenBank the protein coding sequences are found in the features table with the key CDS (CoDing Sequence). The raw sequence, which is the primary data of each GenBank record, may not be anywhere in our sequence set. Hence, our sequences correspond to coding sequence features in GenBank; the GenBank accession and identifiers specify the raw sequence, not specific coding sequences.

Also, in our database, protein sequences are treated as entirely derivative from the nucleotide sequences and are maintained in a separate "translation" table rather than in the main sequence table.

Some complications can arise because of this difference in structure. Because we tie metadata to the influenza *strain* and GenBank maintains multiple copies of this information in and across sequence records, it is possible to receive conflicting information from different records concerning a single strain. When different GenBank entries referring to a particular strain disagree, we need to note the inconsistency and resolve the conflict, possibly through human interaction before updating our database.

3.1.3 Import of GenBank Data

Establishment of a workable automatic process to import GenBank data both incrementally and in bulk required a clear understanding of the existing influenza data and its data quality issues. Some of the key considerations involve:

- Strain identifiers
- Sequence identifiers
- Incremental updates and model maintenance

3.1.3.1 Strain Identification

Each physical sample from which influenza is isolated and sequenced becomes a “strain” according to convention of influenza naming. The names, therefore, when used consistently are sufficient identifiers for influenza strains. Unfortunately, the naming convention is not precisely followed, nor is any format checks performed on historical or contemporary names. For this reason, we have to apply heuristic rules for extracting the proper strain name from the name as given in GenBank. Once established, we use this cleaned up name as the proper way to identify all of the sequences which come from a common source/strain. The database, of course, maps these to integer identifiers which are used internally as proper unambiguous strain keys.

3.1.3.2 Model Maintenance

Because our models are derived from the sequence data, it is important to keep them up-to-date when new information becomes available. Regularly, NCBI publishes updates to the GenBank database that are mostly new sequences, but also updates to existing sequences. Therefore, we must have a mechanism to identify which models are potentially affected by a recent update.

The simple mechanism that we will maintain to manage the model update process is an additional set of database tables that record which sequences are used in each model. The GenBank update process records the sequences that have been changed (using sequence identifiers) and a simple query can indicate which models need to be rebuilt.

3.2 InfluenzO: An Ontology for Influenza

This part of the Bioforensics MSR built the Influenza Ontology (InfluenzO) to connect genomic data with epidemiological data. This is because in the genomics community, the focus is on molecular biology - such as the DNA and protein levels of an ***organism*** - and attempts to link organism properties to variations in the genome. At the same time, in the epidemiology community, the focus is on the process of infection and transmission of disease within a host *population.* The host population provides an indication of how a disease strain associated with a sample came to be at that specific place and time. These disciplines, genomics and epidemiology, and therefore their data, do not typically interact. Bioforensics, in contrast, needs data from both the genomics and epidemiological communities and therefore must bridge the gap between them. InfluenzO links the specific molecular biological properties of a sample from an organism to disease spread within the host population. This is done by formally articulating and encoding explicit definitions of these data and how they relate to each other. This formal representation is called an ontology.

InfluenzO is an “application ontology” (as opposed to a “reference ontology”[3]) and has been designed to be applicable to any collection and sequencing project. This ontology accounts for variables such as event, location, host, outcome, symptoms, pathogenicity and drug resistance. It has been developed in collaboration BioHealthBase, a National Institute of Allergy and Infectious Diseases and National Institute of Health (NIAID/NIH) funded Bioinformatics Research Center (BRC) and with Gemina (Genomic Metadata for Infectious Agents), at the Institute for Genome Sciences at the University of Maryland.

InfluenzO was developed with two use cases in mind: research and surveillance. These use cases enabled us to know what questions to ask of the data. The research use case was created by the Centers for Excellence for Influenza Research and Surveillance (CEIRS) for both ontology development and evaluation. The goal of the research use case was to capture

[3] The Open Biomedical Ontologies (OBO) Foundry classifies ontologies into two types: application and reference. Application ontologies are created to capture relevant fields for a defined purpose. Reference ontologies record how things are naturally in the world and serve as reference for other ontologies.

investigation of data collected on influenza strain mutations that cause death in birds. They chose a specific published project that investigated the effects of PB1-F2 sequence polymorphisms on influenza infection[4] in order to guide the extension of the BioHealthBase database system to support primary CEIRS experiment data and to establish minimum information standards for unambiguous representation of viruses. That particular paper was chosen as a good representation of a typical research paper generated by one of the CEIRS. The ontological components were biomaterial transformations, assays, and data transformations.[5]

The CEIRS surveillance use case is currently under development. Both of these use cases have established or are establishing minimum information standards.

We also developed an internal surveillance use case for this MSR. This surveillance use case considered three clinical samples taken on the same day in Indonesia. GenBank listed three H5N1 samples collected on May 23, 2006 from 10 year old females in Indonesia. One sample was a nasal swab, another was from the pleural fluid, and the third was a throat swab. All GenBank fields for these samples were identical to each other except for the isolation source. Based on examination of these fields, a plausible inference is that these three samples came from the same person. The goal for the ontology was to capture the metadata so that an analyst could readily examine patterns in the data and make appropriate inferences. For example, imagine that samples were taken from a classroom of students who were all the same age. If these samples were entered into GenBank, most of the data fields would be identical to each other. If ten samples were taken, it is not possible, without explicit encoding of unique “patient” identifiers, to determine whether the samples were taken from ten different people or taken from the same person ten times. This example points out the need to capture detailed clinical information, or – in its absence – to bring this to the analyst’s attention, so that s/he can draw a plausible inference based on expert knowledge.

In addition to the internal surveillance use case mentioned above, we identified a second surveillance use case. This scenario would use the ontology to identify strain sequences that demonstrate drug resistance. The ontology would highlight the collected samples that have an amino acid sequence that have been shown to confer drug resistance to the virus.

3.2.1 Building an Ontology

To identify what metadata we needed to include in the ontology identified stakeholders for each of our institution’s projects and compiled them from a collection of terms used in influenza-related projects. As shown in Table 1, numerous sources contributed to terms

[4] Conenello, G.M., Zamarin, D., Perrone, L.A., Tumpey, T. and Palese, P. (2007) A single mutation in the PB1-F2 of H5N1 (HK/97) and 1918 influenza A viruses contributes to increased virulence, *PLoS Pathog*, **3**, 1414-1421.

[5] CEIRS BHB Pilot Project – MSSM PB1-F2 N66S and Virulence. Version 1.1 – 06JAN2007.

including the BioHealthBase BRC6, the Centers for Influenza Research and Surveillance (CEIRS)7, the Gemina project8, and us. In addition, Vik Subbu (Institute for Genome Sciences), contributed terms related to influenza virus samples; three of the six CEIRS (U.C.L.A., Emory, and Mount Sinai Medical Center) contributed surveillance terms; and Florence Bourgeois and Naomi Sengamalay (Children's Hospital Boston) contributed epidemiological terms. Within the context of an Excel spreadsheet, we compiled, aligned, and regularized approximately two hundred terms. Collecting terms was an iterative process that expanded our ontology as we encountered additional collaborators. These terms comprise the metadata about the virus, such as its transmission methods, environmental factors, clinical aspects, and other epidemiological factors. We also included metadata about the institution, the researchers, and the assays used.

Table 1: The Number of Terms Related to Influenza Research and Surveillance that were Contributed to the Initial Influenza Ontology and their Source

Source	No. of Terms
BioHealthBase BRC (U. Texas)	98
CERIS (combined)	83
GEMINA (U. Maryland)	26
HEDDS[9]	48
LEMUR (Children’s Hospital Boston)	94
MITRE (Genomics for Bioforensics MSR)	36

Once the basic alignment was completed, new terms or sets of terms were added to the collection aligned with the existing terms according to their definition. This helped to minimize overlap and duplication of terms. If a term did not map to any of the existing terms we had, it became a new term in the ontology. Once these terms were collected, we assembled related terms into groups and formalized these terms into a controlled vocabulary. Each term was issued a unique identifier.

3.2.2 Collecting Terms

InfluenzO is aligned with the NIH interest in interoperability of data in the biomedical domain by following the Open Biomedical Ontologies (OBO) Foundry development philosophy. Thus, InfluenzO interoperates with other OBO Foundry ontologies. OBO

[6] http://www.biohealthbase.org

[7] http://www3.niaid.nih.gov/research/resources/ceirs/

[8] http://gemina.igs.umaryland.edu

[9] **H**ighly Pathogenic Avian Influenza **E**arly **D**etection **D**ata **S**ystem (HEDDS); http://wildlifedisease.nbii.gov/ai/

Foundry and NIH have divided the biomedical ontology space into non-overlapping and interoperable ontologies. This is where the concept of an “application” ontology comes in; many of the terms in InfluenzO are more appropriately placed in other “reference” ontologies, but are brought together in for an explicit purpose. The ontologies referenced in InfluenzO are listed in Table 2. The benefit of linking with other ontologies is that we could utilize their expertise in their respective fields instead of trying to replicate their efforts. For example, rather than redefining the term “fever” in InfluenzO, we link to it in the Disease Ontology, which is a reference ontology. In addition, InfluenzO is an extension of the infectious disease ontology (IDO) and thus inherits certain terms directly from IDO. For example, the terms pathogenicity, host, vector, and vaccine are from IDO.

Table 2: Ontologies Referenced in InfluenzO

Acronym	Definition
OBI	Ontology of Biomedical Investigations (http://obi-ontology.org/)
EnvO	Environmental Ontology (habitat of pathogen) (http://gensc.org/gc_wiki/index.php/EnvO_Project)
GO	Gene Ontology (http://www.geneontology.org/)
GAZ	Gazetteer (geographic locations) (http://sourceforge.net/projects/obo)
FMA	Foundational Model of Anatomy (http://sig.biostr.washington.edu/projects/fm/)
DC	Dublin Core (http://dublincore.org)
PATO	Phenotype (http://www.bioontology.org/wiki/index.php/PATO:Main_Page)
SO	Sequence Ontology (sequence features) (http://www.sequenceontology.org/)
Cell	Cell Ontology (types of cells) (http://www.obofoundry.org/cgi-bin/detail.cgi?id=cell)
DO	Disease Ontology (http://diseaseontology.sourceforge.net/)
IDO	Infectious Disease Ontology (http://www.infectiousdiseaseontology.org/)

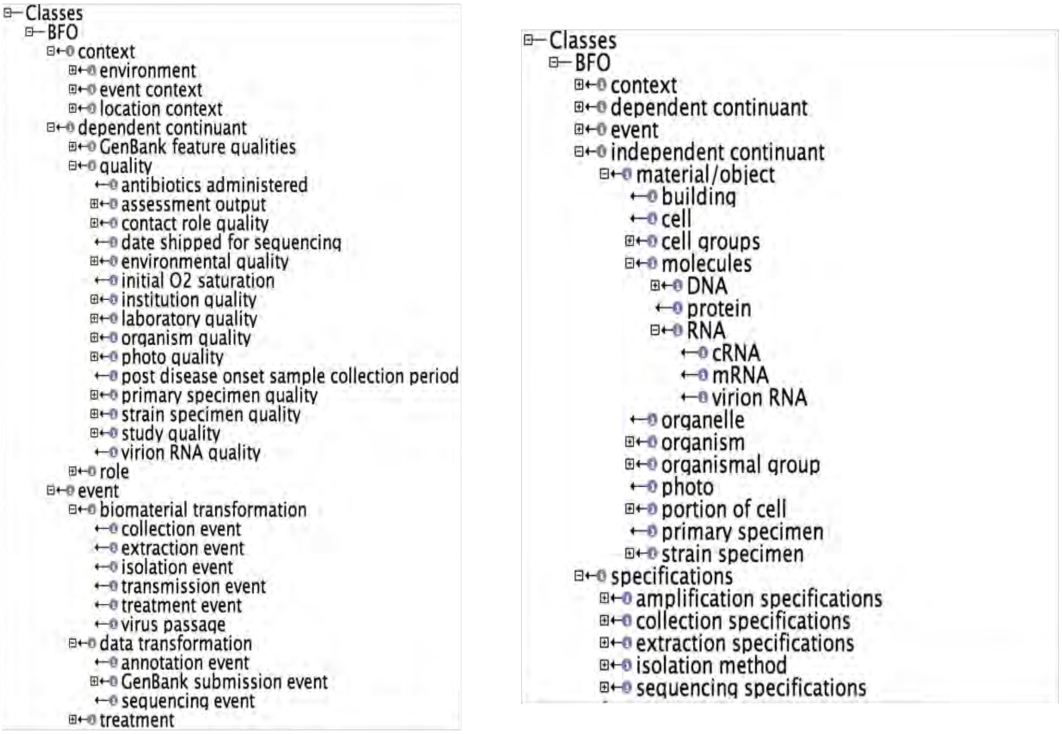
3.2.3 Identifying and Defining Relationships

The purpose of an ontology is to give meaning to the data and place it in context. This is done by identifying what we need to keep track of (the terms or names that identify them) in order to support a use case (influenza research and surveillance) and then formally describe how they relate to one another. For example, we know that both ducks and chickens are birds, but if we do not place them in relation to one another as subtypes (kinds) of birds, the system will not be able to answer a query that asks for all the cases of avian flu in a country in a given year.

3.2.4 Creating a Class Hierarchy

Because the InfluenzO class hierarchy was implemented to align with the OBO Foundry, we used the Basic Formal-Ontology (BFO) class hierarchy structure. This structure was based on two distinctions drawn from the philosophical literature a) *Continuant* and *Occurrent* entities, and b) *Dependent* and *Independent* entities. Continuants are entities which continue to exist through time. Occurrents unfold themselves through time, in successive temporal phases. An *Independent* entity asserts that it has an inherent ability to exist without reference to other entities.

Underneath the BFO "upper ontology" (Figure 2, right) we added the independent continuant classes for Primary Specimen, Amplified Strain Specimen and Vaccine. Dependent continuants capture the qualities of the above (Figure 2, left).



- Classes
  - BFO
    - context
      - environment
      - event context
      - location context
    - dependent continuant
      - GenBank feature qualities
      - quality
        - antibiotics administered
        - assessment output
        - contact role quality
        - date shipped for sequencing
        - environmental quality
        - initial O2 saturation
        - institution quality
        - laboratory quality
        - organism quality
        - photo quality
        - post disease onset sample collection period
        - primary specimen quality
        - strain specimen quality
        - study quality
        - virion RNA quality
      - role
    - event
      - biomaterial transformation
        - collection event
        - extraction event
        - isolation event
        - transmission event
        - treatment event
        - virus passage
      - data transformation
        - annotation event
        - GenBank submission event
        - sequencing event
      - treatment

- Classes
  - BFO
    - context
    - dependent continuant
    - event
    - independent continuant
      - material/object
        - building
        - cell
        - cell groups
        - molecules
          - DNA
          - protein
          - RNA
            - cRNA
            - mRNA
            - virion RNA
        - organelle
        - organism
        - organismal group
        - photo
        - portion of cell
        - primary specimen
        - strain specimen
    - specifications
      - amplification specifications
      - collection specifications
      - extraction specifications
      - isolation method
      - sequencing specifications

Figure 2: InfluenzO Classes. The Dependent Continuants (left) Capture the Qualities of the Independent Continuants (right)

The ontology is formalized in the OBO language using OBO-Edit[10], a freely available tool from the Gene Ontology consortium.

[10] http://www.oboedit.org

3.2.5 Evaluation

The CEIRS research use case (described above) was to serve as an evaluation use case for the Influenza Ontology using an evaluation approach that was developed under a separate funding (an Ontology Evaluation IG). The main aspect of the approach, which was aimed at evaluating an ontology for reuse, was that once these experiment/protocol applications are encoded into ontological "semantic components" they can be independently validated and reused. That meant that research papers encoded to test capture of the data, methods and results would enable further investigation of influenza strain mutations and their virulence. The evaluation criteria metrics considered are correctness, how well the science is captured, completeness, percent coverage of terms, and utility of the ontology. Percent coverage of terms reflects whether the terms of the minimum information standard from the use case were captured. The utility would be assessed by a series of "competency" or "challenge" questions.

The InfluenzO was shared with the community at the IDO Workshop in September 2008 where BioHealthBase presented the development work. Workshop attendees stressed the community's need for the InfluenzO and requested additional considerations, such as virulence definitions and ontology evaluation in general. It is anticipated that a paper will be published on the InfluenzO's development. Additional future work includes formalizing and validating the ontology using CEIRS and BioHealthBase use cases, circulating the ontology through the community for its review, and integrating it into the workflow.

3.2.6 Current State of Influenza Ontology

The pre-release of InfluenzO, version 0.17 is available on SourceForge[11], at the time of this report. Once the quality control checks are completed and the database references filled in, the first release of InfluenzO, version 1.0, will be made available on the project website at SourceForge. At the same time, it will be submitted to submit it to the OBO Foundry. Updates of on the development of InfluenzO are maintained on the InfluenzO Wiki[12].

In this process, we have learned some important lessons learned that will affect our and, more importantly, the public's ability to create an environment to support infectious disease research and surveillance. The most relevant is an urgent need for data standards (structured vocabularies, ontologies, tools for data collection). Currently, genomics sequence data is collected and stored separately from most of the epidemiological data. This means that critical metadata – such as host demographic information, clinical manifestations, and outcome – is lost – for example, did the host die or survive? In addition, metadata about sample collection location, sample collection procedures, and sequencing techniques are recorded only very sporadically. To improve this, we have developed and made available

[11] http://sourceforge.net/projects/influenzo/
[12] http://influenzaontologywiki.igs.umaryland.edu

software to parse the influenza strain name for geospatial metadata, to enhance specificity of geographic metadata.

We hope that capture of metadata will be improved through the introduction of structured vocabularies and ontologies, coupled with tools to make data capture and data submission processes easy to use. This work ties into the effort by the CEIRS to focus on data interoperability and data standards (e.g., InfluenzO) as well as the work of the Genome Standards Consortium[13] to provide better standards for the collection and the structuring of metadata.

[13] http://gensc.org/gsc/

4 Automated Genotyping of Influenza Sequences

In the event of a biological incident, whether natural or introduced, the ability to quickly characterize a strain based on its sequence is of great use. The classic approach to this is to produce a phylogenetic tree, starting with a multiple sequence alignment. However, these methods are impractical for rapidly growing datasets such as influenza (for example, the CEIRS will support rapid sequencing of samples during outbreaks), due both to the increase of computational complexity as the sequence count increases, as well as to the significant human effort necessary to interpret the resulting tree. An alternative to these methods are computational genotyping methods, which group a set of sequences into partitions based on their sequences. Under this MSR, we have developed a set of tools for quickly genotyping a set of DNA or protein sequences

4.1 Computational Genotyping with Complete Composition Vectors and Affinity Propagation Clustering

As noted above, sequence comparison traditionally begins by building a multiple sequence alignment (MSA) of the sequences that are being compared. Because of the fact that the structure of the alignment is dependent on the entire sequence set, whenever a sequence is added to the dataset, the MSA must be recomputed. As the number of sequences increases, there is a tradeoff between computation time and accuracy – many of the most accurate algorithms cannot be used on large datasets.

Recently, Wu and colleagues[14] proposed a novel method for comparing genetic sequences. This method, known as the Complete Composition Vector (CCV) method, uses sliding windows of varying lengths to describe each sequence as a numerical vector. The distances between these vectors can be calculated using any standard distance metric (e.g., Euclidian, cosine, Manhattan). The main advantage here is that the pairwise distances are no longer dependent on an entire set of sequences – as sequences are added to the dataset, the distance matrix only need be expanded, not completely recalculated. We demonstrated that the CCV method, combined with a cosine distance metric, produced trees with nearly the same topology as those produced with traditional methods[15].

In order to genotype the sequences, we applied a recently developed clustering algorithm called affinity propagation clustering[16] to the distance matrix produced by the complete composition vector methods. In order to test these methods, we genotyped a set of 94 H5 HA sequences that had been manually curated by the WHO and OIE as part of a movement to

[14] Wu, X., Wan, X.-F., Wu, G., Xu, D. and Lin, G. (2005) Whole Genome Phylogeny via Complete Composition Vectors. *Technical Report TR05-06*. Department of Computing Science, University of Alberta.

[15] Peterson, MW, Mardis S, Colosimo M, and Hirschman L (2008) Automated Genotyping via Complete Composition Vectors and Affinity Propagation Clustering. Submitted to *Bioinformatics.*

[16] Frey, B.J. and Dueck, D. (2007) Clustering by passing messages between data points, *Science*, **315**, 972-976.

generate standard nomenclature set[17] for H5N1 Avian influenza. Our method produced clusters that correlated well with the expertly-defined clades.

In order to test the utility of this system for microbial forensic analysis, we recreated a scenario in which sequence analysis was used to isolate the source of an influenza outbreak. In February 2007, an outbreak of H5N1 Avian influenza A was reported at a turkey farm in Suffolk, England, killing over 2000 birds, and leading to the cull of over 150,000 more. This raised red flags since there had not been a reported case of H5N1 in the United Kingdom since April 2006. A week after the outbreak, a government veterinarian noted that the owner of the farm also owned a poultry packing plant in Hungary, where there had been an outbreak of H5N1 two weeks earlier. Upon sequencing the UK sequence and comparing them to samples taken from infected geese in southeastern Hungary, they were found to be 99.6% similar. Government epidemiologists concluded that the most likely cause of the outbreak was the import of poultry from the plant in Hungary, from where over 100 tons of poultry were brought to the packing plant adjacent to the farm in England during the outbreak.

We recreated this scenario by taking the sample representing the UK outbreak and creating a background dataset of HA sequences representing all similar strains in the database (see following section for more detail). The CCV-affinity propagation genotyping pipeline was applied to this dataset, resulting in nine clusters. A phylogenetic tree, with the leaves colored by cluster membership, can be seen in Figure 3. The UK samples were found to form a tight cluster with two samples from Hungary, representing the earlier outbreak (seen in red box). Our methods arrived at the same conclusion as traditional methods, with minimal human intervention.

A manuscript describing this work is being prepared for submission.

[17] http://www.who.int/csr/disease/avian_influenza/guidelines/nomenclature/

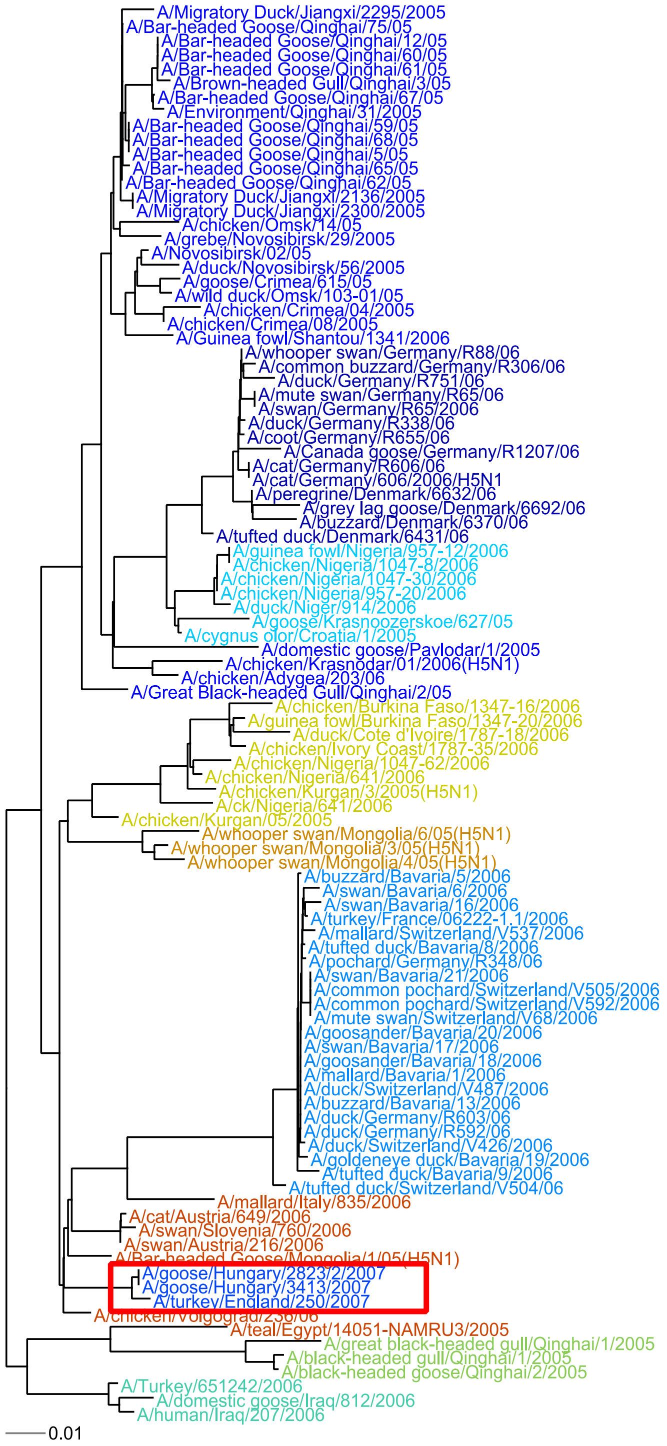

A/Migratory Duck/Jiangxi/2295/2005
A/Bar-headed Goose/Qinghai/75/05
A/Bar-headed Goose/Qinghai/12/05
A/Bar-headed Goose/Qinghai/60/05
A/Bar-headed Goose/Qinghai/61/05
A/Brown-headed Gull/Qinghai/3/05
A/Bar-headed Goose/Qinghai/67/05
A/Environment/Qinghai/31/2005
A/Bar-headed Goose/Qinghai/59/05
A/Bar-headed Goose/Qinghai/68/05
A/Bar-headed Goose/Qinghai/5/05
A/Bar-headed Goose/Qinghai/65/05
A/Bar-headed Goose/Qinghai/62/05
A/Migratory Duck/Jiangxi/2136/2005
A/Migratory Duck/Jiangxi/2300/2005
A/chicken/Omsk/14/05
A/grebe/Novosibirsk/29/2005
A/Novosibirsk/02/05
A/duck/Novosibirsk/56/2005
A/goose/Crimea/615/05
A/wild duck/Omsk/103-01/05
A/chicken/Crimea/04/2005
A/chicken/Crimea/08/2005
A/Guinea fowl/Shantou/1341/2006
A/whooper swan/Germany/R88/06
A/common buzzard/Germany/R306/06
A/duck/Germany/R751/06
A/mute swan/Germany/R65/06
A/swan/Germany/R65/2006
A/duck/Germany/R338/06
A/coot/Germany/R655/06
A/Canada goose/Germany/R1207/06
A/cat/Germany/R606/06
A/cat/Germany/606/2006/H5N1
A/peregrine/Denmark/6632/06
A/grey lag goose/Denmark/6692/06
A/buzzard/Denmark/6370/06
A/tufted duck/Denmark/6431/06
A/guinea fowl/Nigeria/957-12/2006
A/chicken/Nigeria/1047-8/2006
A/chicken/Nigeria/1047-30/2006
A/chicken/Nigeria/957-20/2006
A/duck/Niger/914/2006
A/goose/Krasnoozerskoe/627/05
A/cygnus olor/Croatia/1/2005
A/domestic goose/Pavlodar/1/2005
A/chicken/Krasnodar/01/2006(H5N1)
A/chicken/Adygea/203/06
A/Great Black-headed Gull/Qinghai/2/05
A/chicken/Burkina Faso/1347-16/2006
A/guinea fowl/Burkina Faso/1347-20/2006
A/duck/Cote d'Ivoire/1787-18/2006
A/chicken/Ivory Coast/1787-35/2006
A/chicken/Nigeria/1047-62/2006
A/chicken/Nigeria/641/2006
A/chicken/Kurgan/3/2005(H5N1)
A/ck/Nigeria/641/2006
A/chicken/Kurgan/05/2005
A/whooper swan/Mongolia/6/05(H5N1)
A/whooper swan/Mongolia/3/05(H5N1)
A/whooper swan/Mongolia/4/05(H5N1)
A/buzzard/Bavaria/5/2006
A/swan/Bavaria/6/2006
A/swan/Bavaria/16/2006
A/turkey/France/06222-1.1/2006
A/mallard/Switzerland/V537/2006
A/tufted duck/Bavaria/8/2006
A/pochard/Germany/R348/06
A/swan/Bavaria/21/2006
A/common pochard/Switzerland/V505/2006
A/common pochard/Switzerland/V592/2006
A/mute swan/Switzerland/V68/2006
A/goosander/Bavaria/20/2006
A/swan/Bavaria/17/2006
A/goosander/Bavaria/18/2006
A/mallard/Bavaria/1/2006
A/duck/Switzerland/V487/2006
A/buzzard/Bavaria/13/2006
A/duck/Germany/R603/06
A/duck/Germany/R592/06
A/duck/Switzerland/V426/2006
A/goldeneye duck/Bavaria/19/2006
A/tufted duck/Bavaria/9/2006
A/tufted duck/Switzerland/V504/06
A/mallard/Italy/835/2006
A/cat/Austria/649/2006
A/swan/Slovenia/760/2006
A/swan/Austria/216/2006
A/Bar-headed Goose/Mongolia/1/05(H5N1)
A/goose/Hungary/2823/2/2007
A/goose/Hungary/3413/2007
A/turkey/England/250/2007
A/chicken/Volgograd/236/06
A/teal/Egypt/14051-NAMRU3/2005
A/great black-headed gull/Qinghai/1/2005
A/black-headed gull/Qinghai/1/2005
A/black-headed goose/Qinghai/2/2005
A/Turkey/651242/2006
A/domestic goose/Iraq/812/2006
A/human/Iraq/207/2006
0.01

Figure 3: Genotyping for Forensic Analysis

4.2 Automated Background Selection with BLAST

The influenza analysis workbench allows the user to compare a given, presumably novel, sample against a background of comparable sequences from the same organism. Inferences can be drawn about the novel sample by considering how near or far it is from samples (and clusters of samples) in the background model, considering both genetic distance and similarity of metadata (geospatial location, time). Using the entire collection of GenBank influenza A data is not practical because of it present and future size.

4.2.1 Need for Smaller Models for User Comparison

Even it were computationally feasible to make background models with hundreds or thousands of sequences, it is beyond the capability of our visualization techniques to represent the results in a way comprehensible to an investigator. Therefore, we also wish to reduce the size of background models to that which can be adequately displayed and interpreted. With our present set of techniques, a generous upper bound on the number of sequences would be around 100, though many of the trees start to become dense and confusing at that size. Though we have not performed studies to refine the number, a practical background size would be 30-60 sequences.

4.2.2 Sampling Inconsistency

One factor complicating the selection of background models is the inconsistency of sampling. As sequencing technology has advanced, the number of sequences available has increased dramatically giving a much more detailed view of recent outbreaks and fairly limited coverage of older ones. The variability of sampling over geographic location is a larger concern. With the exception of large global surveillance efforts in anticipation of the spread of avian influenza, collection of influenza samples is sporadic, with relatively few sequences taken from the poorer parts of the world. A further complication is that occasionally a sample source will get sequenced multiple times. This can happen if multiple samples are drawn from the same source either over time or in different ways (blood vs. sinus). With influenza, these various samples nearly always remain identical though they are represented by multiple GenBank records.

The net effect of this is that if one looks at a neighborhood of sequences in the vicinity of any particular sequence, that neighborhood can represent a very narrow range of time and place for some sequences and can cover many years and most of the globe for others. If neighborhoods are instead defined by a reasonable genetic distance (97% similarity), the result is a few very large clusters including hundreds of similar samples and a few hundred very small clusters with a few samples (2-5).

4.2.3 Sequence Replication

One simple approach to reducing the size of the models without sacrificing coverage would be to eliminate sequences that are exact duplicates of others in the database. As mentioned previously, a sample will get occasionally be sequenced multiple times and is therefore represented by multiple GenBank records. These could certainly be eliminated. More difficult is attempting to decide when sequences that differ slightly can be reduced in number, perhaps represented by a single sequence. Thinning the data in this way depends, of course, on what is needed for the attribution task. Sequences that are very similar genetically but were taken from hosts in very different locations might be essential to understanding the background prevalence of certain strains of influenza and therefore vital to assigning a probable source to a novel sample. Only when a sequence is nearly identical in both its genetic and metadata characteristics can it be safely removed from a background model without compromising the completeness of the comparison.

At present, we have not implemented any data set technique aimed at reducing the number of less informative sequences. Our approach thus far has been to create reasonably sized neighborhoods regardless of how much near duplication is present.

4.2.4 Approximate Clustering using BLAST

The approach we have chosen to group sequences for the creation of background models is to use BLAST (the Basic Local Alignment Search Tool) to create clusters of genetically similar sequences. Our approach is motivated the canopy clustering algorithm introduced by McCallum *et al*.[18].

The algorithm is suited to performing clustering on a very large number of data points where the clustering cost is fairly high. The data is divided into partially overlapping subsets using a distance metric that is cheap to produce; clustering is performed on each of the subsets, and then the cluster results are combined into a composite clustering of all the data. In our application, we do not need (at present) to perform the composite clustering but we will use BLAST as a pair-wise approximation of the genetic distance calculated by our CCV approach. As in the canopy approach we create overlapping neighborhoods so that meaningful clusters of sequences do not get placed in separate neighborhoods.

BLAST, which functions very much like an information retrieval tool for genetic sequences, returns several similarity scores that can be used to cluster sequences. It also ranks results in order of similarity, allowing a program to easily collect the nearest neighbors to any given sequence. Our approach was to create neighborhoods by choosing a sequence at random, running a BLAST query, and collecting nearby sequences using a similarity threshold. Using a tighter threshold we eliminate sequences from our base set, effectively allowing sequences between thresholds to appear in the overlap between neighborhoods. We then iterate, using other randomly chosen sequences as the neighborhood seed, until all sequences have been assigned.

[18] McCallum A, Nigam K and Ungar L (2000) L.H.: Efficient clustering of high-dimensional data sets with application to reference matching. Knowledge Discovery and Data Mining. 169-178.

4.2.5 Sequence Incompleteness and Metadata Errors

In preparing our experiments we identified several inconsistencies in the GenBank records that complicated the task. First of there are a number of sequences that are simply mislabeled with the wrong gene product. For example there are a few sequences for producing the influenza nonstructural 2 protein (NS2) that are identified as nonstructural 1 (NS1), and vice versa. Similarly there are hemaglutinin (HA) sequences that record only the portion of the gene for the HA1 domain even though they are marked as HA. Still others record only a fraction of the entire protein coding sequence, without any indication of how long the complete sequence would be, had it been fully sequenced. Each of these complicates the task of identifying all the sequences that can be effectively used in the context of the workbench. The varying length and partial subsequences further affect the scoring we are using to establish neighborhood thresholds.

Another complication that we were required to accommodate was that the BLAST scores do not constitute a valid metric, i.e., they are not symmetric. This is largely an artifact of comparing sequences of differing lengths; BLAST scores are calibrated to the length of the query sequence. In order to make the scores suitable for selecting clusters, we renormalize them as if the retrieved sequences were as long as the query sequence. This is based on an assumption that differences in nucleotides between two sequences are uniformly distributed across the sequence. That is, the probability of mutation of any nucleotide is the same everywhere. With sequences of roughly similar length, this assumption is probably reasonably robust. For sequences of drastically different lengths, this certainly fails and for this reason we drop from consideration any retrieved sequence that is not at least 80% as long as the query sequence.

4.2.6 Selecting Among Background Models

In the final attribution system it is necessary to select which of our potentially numerous background models are appropriate for comparison against a novel sample. The most straightforward way to select among background models is to use BLAST again. Since BLAST can efficiently find the nearest sequences among all models, it is a relatively simple matter to choose a background model using the best matched sequences from a BLAST query. It is conceivable that the separate segments of the influenza genome will not point to the same background models. In these cases it will be necessary to prepare a display with results for multiple models along with a clear indication of the possibility of a reassortment event having taken place. That said, the background models should not be so tightly focused as to cause reassortments to frequently span models.

A similar and possibly tractable alternative to model selection would be to judiciously reuse the CCV framework for selecting among models. While the total number of sequences prohibits using the framework as is for identifying appropriate models, there are ways in which we can reduce the number of sequences necessary for a "cross-model" model. One approach would be to subsample a model for sequences that are "typical" of the group, perhaps by identifying sequences whose variance with other sequences in the model is minimal. Though time consuming, a second approach would be to calculate a consensus sequence of the model after performing a multiple sequence alignment. Each model would then be represented by a single sequence in a higher-level CCV model. The most direct approach would be to use the features of the CCV itself, selecting out features (k-mers) that are most consistent across the model. This is just the opposite of the step used in calculating CCV distances where the most highly discriminatory features are chosen. Once a large set of features that are consistent across a model are identified, these features can be used in comparison to other models to identify features that are discriminatory between models, rather than within them.

4.3 Reassortment Detection

Because influenza is a segmented genome, multiple strains of the virus can infect the host, mix their representative segments, and create a novel strain. This process, known as reassortment, lead to the 1957 “Asian Flu” and the 1968 “Hong Kong” pandemics, and it is a likely mechanism for the next pandemic strain. In addition, the presence of reassortment hinders attribution – the most similar strains will change depending on which gene is examined.

To test for reassortment between two genes, we normalized the distance matrices for each gene so that the largest distance between two samples is 1, and the smallest is 0. The absolute distance between each element in the two matrices is calculated, and this matrix is clustered using by complete linkage clustering. The assumption here is that samples which are reassorted exhibit large differences between genes, while non-reassortant samples have a distance near zero. The separation of the two largest clusters was used as a measure of confidence that reassortment had occurred.

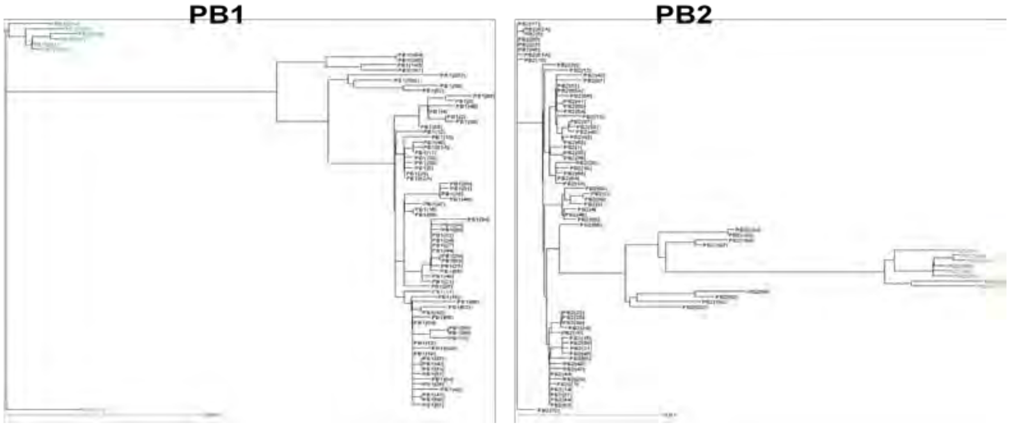


Figure 4: Reassortment Event between PB1 and PB2 Genes

We tested our method on a set of sequences collected by The Institute for Genome Research (now the J. Craig Venter Institute) between 1999-2004 and studied extensively for reassortment events[19]. This set of sequences was shown to have been involved in many reassortment evens. One of these events, between the PB1 and PB2 genes, is represented by a single sample (A/New York/11/2004). This event is shown in the trees in Figure 4. These two genes cluster into two main groups, represented in green and black. A/New York/11/2004 (colored in red) clusters with the larger (black) clade in the PB1 gene, while it clusters with the smaller (green) clade PB2.

We ran our test for reassortment on the PB1 and PB2 datasets, and found a cluster separation of 0.9067, indicating a high likelihood of reassortment. One of the two clusters was comprised of a single sample: A/New York/11/2004, the reassortant sample. We were also able to identify other, larger, known reassortment events with our test.

[19] Holmes, E.C., Ghedin, E., Miller, N., Taylor, J., Bao, Y., St George, K., Grenfell, B.T., Salzberg, S.L., Fraser, C.M., Lipman, D.J. and Taubenberger, J.K. (2005) Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses, *PLoS Biol*, **3**, e300.

5 Visualization of Sequence Data and Metadata

5.1 TreeViewJ

As discussed in the previous section, traditional comparative genomics work is done via the examination of a phylogenetic tree, which represents the (usually inferred) evolutionary relationship between a set of sequences thought to descend from a common ancestor. While these trees are invaluable for comparative genomic analysis, any questions researchers may have are often answered using the metadata associated with the primary sequence data. One of the more important pieces of metadata for a microbial forensics or analysis is the time of sample collection.

Since none of the available tree visualization tools met our needs, we developed our own phylogenetic visualization tool, TreeViewJ. This tool reads and writes phylogenetic trees in a variety of formats, including Nexus, Newick, and PhyloXML. PhyloXML is a recently developed XML format for describing phylogenetic trees. Trees can be visualized as cladograms (trees showing the inferred evolutionary relationships, but not distances) or phylograms (trees which incorporate distance measurements). The software allows the user to color and label the tree, and save to a JPG or SVG format to produce images for publication.

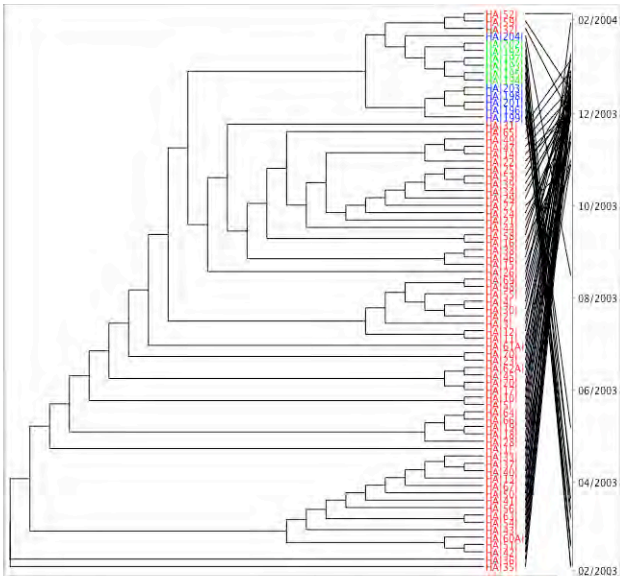


Figure 5: TreeViewJ Timeline View

In addition to the visualization options, TreeViewJ allows the user to map a date to each leaf of the tree. Once this is done, the user can sort the tree by date, resulting in a tree which, while isomorphic to the original tree, is sorted in either ascending (youngest sample on top) or descending (oldest sample at top) order. A timeline display is also available, in which a line is drawn from each leaf of the tree to its corresponding date on the timeline. This view can be seen in Figure 5.

TreeViewJ has been released under the Gnu Public License, and is available at http://treeviewj.sourceforge.net. In addition, a paper describing the publication has been published in Source Code for Biology and Medicine[20].

5.2 Visualization of Genotyping Results using Web 2.0 tools

By genotyping the sequences into discrete groups, rather than solely producing a phylogenetic tree for genetic analysis, we are able to incorporate a measure of genetic similarity (genotype membership) with the sequence metadata (geographic location, time of collection, host information, etc.) By combining this information in an intelligent way, we aimed to enable analysts to make decisions based on these data quickly and confidently. The workbench was developed as a web application to avoid potential problems with compatibility across different target platforms. The Dojo Toolkit[21] was used as a foundation. This toolkit provides the necessary user interface (UI) tools for the development of full-featured applications which run inside a web browser. Figure 6 shows a Dojo window with a bar chart of the host information for a query. The different colors represent each genotype found in the dataset.

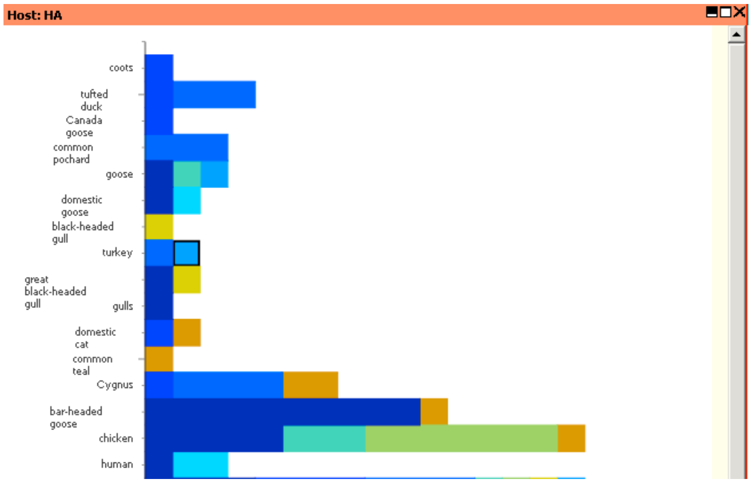
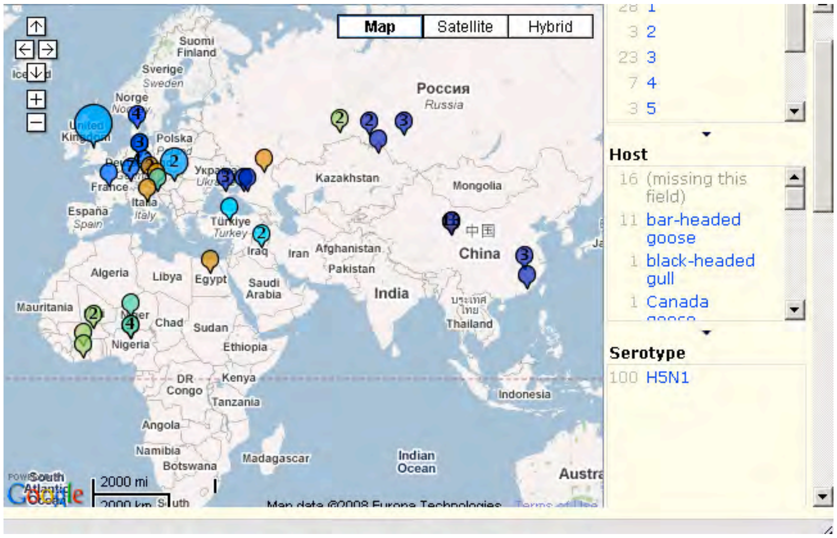


Figure 6: Dojo Window and Chart

[20] Peterson, M.W. and Colosimo, M.E. (2007) TreeViewJ: an application for viewing and analyzing phylogenetic trees, *Source Code Biol Med*, **2**, 7.
[21] http://dojotoolkit.org

For geospatial and temporal information, the Exhibit toolkit, part of the SIMILE project[22] was used. Exhibit provides the ability to apply a faceted browsing approach to the visualization of the data – the user can filter the data based on a certain feature (e.g. genotype, host, serotype), and all of the views are updated to only show the samples which match that criteria. Figure 7 shows Exhibit's map view inside a Dojo window. The query sample is identified by the largest marker, and the most similar samples (members of the same genotype) are identified by their color and the intermediate marker size. The dissimilar samples are colored by their cluster membership, and are marked using a smaller marker.



28	1
3	2
23	3
7	4
3	5

Host

16	(missing this field)
11	bar-headed goose
1	black-headed gull
1	Canada goose

Serotype

100	H5N1

Figure 7: Exhibit Map View

Together, these tools provide an extensible, easy-to-use system for combining sequence similarity with sample metadata, and allow the user to make an informed decision based on this information.

[22] http://simile.mit.edu/exhibit/

6 Lessons Learned and Impact

6.1 Lessons Learned

There are a number of important lessons learned that will affect our and, more importantly, the Government's ability to create an environment to support microbial forensics. We list these below.

6.1.1 Urgent Need for Data Standards (Structured Vocabularies, Ontologies, Tools for Data Collection)

Genomics sequence data is collected and stored separately from any kind of epidemiological data. This means that critical metadata – such as host demographic information, clinical manifestations, and outcome – is lost, e.g., whether the host died or recovered. In addition, metadata about sample collection location, sample collection procedures, and sequencing techniques are recorded only very sporadically. We hope that this will be improved through the introduction of structured vocabularies and ontologies, coupled with tools to make data capture and data submission processes easy to use. This ties in to the effort by the CEIRS to focus on data interoperability and data standards (e.g., the InfluenzO) and the work of the Genome Standards Consortium to provide better standards for the collection and structuring of metadata.

6.1.2 Construction of a Representative Reference Database

Under the Genomics for Bioforensics project, we have made significant progress in understanding the issues in construction of a reference database for a disease modeling, epidemiological, or forensics application. Creation of the reference database is complicated by a number of factors:

- Lack of metadata standards and inconsistent collection of metadata.
- Inconsistencies in GenBank itself: multiple sequences for the same gene or genome.
- Inconsistent strain nomenclature: this is a problem for influenza strains, but a much greater problem for microbes, which are often only partially sequenced.
- Quality of sequence data: for most of the sequences, we do not know what the sequencing coverage is or what method was used (this includes number of passages). This is highly important for the interpretations of the distances between strains.
- Quantity and distribution of data: a truly useful reference database for forensics would need representative samples from the entire range of an organism. Again, this may be improving somewhat as various groups focus on large scale pathogen sequencing efforts.

- Currency of data: novel sequencing methods are leading to a data explosion. A reference database will need to support regular updates and have a scalable architecture that will allow it expand exponentially over time.

Under our work on the MSR, we have identified these issues and have put in place a prototype architecture to support a scalable, maintainable reference database of sequence data and associated metadata.

6.1.3 New Phylogenetics Methods for Large Sets of Highly Similar Sequence Data

The standard applications for phylogenetics have been focused on comparing evolution of organisms over millions of years. These methods were designed to handle small sets of widely divergent sequence data. They do not apply well to the forensics application where there is a need to compare and cluster (genotype) growing sets of organisms that may have diverged only slightly over a period of days, months or years. The need for a fast, scalable genotyping system drove much of our research over the past eighteen months, resulting in a novel approach to genotyping and several new tools.

6.1.4 Scalable Multi-Dimensional Visualization of Sequence Data

Currently the main way to identify genotypes is to visualize the strains in a phylogenetic tree and locate and mark the clusters by hand. This method is unscalable, unquantitative, and often uninformative. This is particularly true when trying to understand genome-scale rearrangements (e.g., reassortments in influenza or recombination in microbes). Looking at phylogenetic trees for two genes within the same set of samples will typically produce very different phylogenetic trees. Without better tools, it is difficult to tell whether the trees differ significantly due to a reassortment, where one gene has a different evolutionary history than another gene – or whether this is just statistical variability due to the method of creating and drawing phylogenetic trees. The clustering genotyping algorithm developed under this project provides a good first step, but more research is needed to automate both detection and visualization of reassortment. This is important, since reassortment of influenza strains may lead to new strains that can escape protection provided by the current vaccines.

6.2 Impact

The impact of this work falls into three categories: new data standards for metadata capture, new algorithms, and application to new areas.

6.2.1 New Data Standards

The creation of the InfluenzO has been a collaborative project with two other groups (BioHealthBase and Gemina). This collaboration has provided us access to a much larger research community funded by NIAID, namely the CEIRS. In particular, our collaborators from BioHealthBase are developing the repository for the influenza data integration activities for the CEIRS. This will provide an application of the InfluenzO for data integration, as well as a potential technology transfer opportunity for the MITRE metadata integration activities, including the geospatial data extraction and display capabilities associated with the Microbial Forensics Workbench, as well as the genotyping software. Through its work on the InfluenzO, MITRE has also been participating in the larger Infectious Disease Ontology working. These activities provide MITRE with a “seat at the table” in defining new data standards and controlled vocabularies/ontologies that will be central to MITRE’s core Biosecurity mission. Dr. John Brownstein (Children’s Hospital Boston) has expressed interest in incorporating the influenza ontology into Healthmap23, a global disease alert system.

6.2.2 New Algorithms

In the “Lessons Learned” section, we noted that we need new genotyping tools, new phylogenetics algorithms and new visualization methods to address the microbial forensics application. This project has made significant contributions to the larger research community in bioforensics and the study of the evolution pathogens. One tool (TreeViewJ) has been published, and the second tool for genotyping has been submitted for publication.

6.2.3 Application to New Areas

One of the most exciting developments has been the application of the general phylogenetics, clustering and visualization tools to new areas and to more complex genomes. There are a number of collaborators who are now interested in applying our tools to their particular problems.

6.2.4 Publications and Posters

- Peterson, M., Mardis, S., Colosimo, M., and L. Hirschman. “Automated genotyping via complete composition vectors and affinity propagation clustering.” In preparation.
- Squires, R., Scheuermann, R., Schriml, L., Bortz, E., Staab, T., Colosimo, M., and J. Luciano. Utilizing the Ontology of Biomedical Investigations (OBI) for Influenza Sequence and Surveillance Analysis, Poster presented at ISMB 2008, Toronto, Canada.
- Luciano, J. Schriml, L., Squires, B. and R. Scheuermann. The Influenza Infectious Disease Ontology (I-IDO). (http://www.bio-ontologies.org.uk/download/Bio-Ontologies2008.pdf) The 11th Annual Bio-Ontologies Meeting, ISMB 2008, Toronto, Canada.
- Morgan, A.A., Lu, Z., Wang, X. Cohen, A.M., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H-H., Torres, R., Krauthammer, M.,

[23] http://www.healthmap.org/en

Lau, W.W., Liu, H., Hsu, C-N., Schuemie, M., Cohen, K.B., and L. Hirschman, Overview of BioCreative II Gene Normalization, Genome Biology 2008, 9(Suppl 2).

- Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., Hirschman, L., and A. Valencia. Evaluation of text mining systems for Biology: overview of the Second BioCreative community challenge, Genome Biology 2008, 9(Suppl 2).
- Krallinger, M., Valencia, A. and Hirschman, L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology, Genome Biology 2008, 9(Suppl 2).
- Altman, R., Bergman C.M, Blake, J, Blaschke, C., Cohen, A., Gannon, F., Grivell, L., Hahn, U., Hersh, W., Hirschman, L., Jensen L.J., Krallinger, M., Mons, B., O'Donoghue, S.I., Peitsch, M., Rebholz-Schuhmann, D., Shatkay, H., and A. Valencia. Text mining for biology - the way forward: opinions from leading scientists, Genome Biology 2008, 9(Suppl 2).
- Matthew Peterson and Marc Colosimo. "TreeViewJ: An application for viewing and analyzing phylogenetic trees." Source Code for Biology and Medicine, October 2007.
- James Dunyak, Marc Colosimo, and Lynette Hirschman. "Poisson Stars and Phylogenetic Trees: Limits of Inferences on Population History," Proceedings of the 2006 Joint Statistical Meetings.
- Field et al. (including Hirschman as signatory), Response to "Putting Molecules on the Map" Nature 453, 978 (19 June 2008) | doi:10.1038/453978b; Published online 18 June 2008.

6.2.5 Presentations and Tutorials

- Joanne Luciano – invited speaker, Disease Ontology Workshop at the European Bioinformatics Institute (EBI)'s Industry Programme, June 2008.
- Joanne Luciano – presenter, Infectious Disease Ontology workshop, Buffalo, New York, September 2008.
- Lynette Hirschman – invited speaker, University of Chicago. "Capturing Context: "Data and Metadata for Metagenomics" April 22, 2008.
- Lynette Hirschman – invited speaker, Manchester University. "Building biological databases: Where can text mining help?" March 18, 2008.
- Lynette Hirschman – invited presenter, EnvO Workshop, Manchester, UK, June 2008.
- Lynette Hirschman – invited presenter, EnvO Workshop, Cold Spring Harbor, NY, November 2007.

- Lynette Hirschman – keynote speaker, Fraunhofer Text Mining Symposium, September 2007.
- Lynette Hirschman – invited speaker, Max Plank Institute of Marine Biology.
- Jim Dunyak – Presenter, 2006 Joint Statistical Meeting, Seattle, WA.
- Jennifer Mathieu and Marc Colosimo – participants in Strong Angel III, San Diego, 2007.

Appendix A Sequence Identifiers in GenBank

Several resources for understanding the intended and actual use of the various GenBank sequence identifiers are found in a number of places, all presently available on the web. Release notes for GenBank[24],[25] describe the basic process for Accession Number (variously called an "Accessions" or "Accession Identifiers"), which are common to GenBank, EMBL, and DDJB, as well as the GenBank specific GI (GenInfo Identifier) numbers used for sequence identification. Release 111.0 notes describe the scheme at the point where Accession and Version numbers were standardized across NCBI, EMBL, and DDBJ. Release 167.0 notes give a more recent update of the agreed upon process.

A much more detailed, though less accessible account of identifier operations is given in the GenBank SDK documentation, under the topic of SEQLOC (sequence locations)[26]. The annotated GenBank sample record[27][4] states with more certainty some of the constraints on identifiers that are equivocally described in other sources.

[24] Distribution Release Notes, Genetic Sequence Data Bank release 111.0, 15 April 1999, ftp://ftp.ncbi.nih.gov/genbank/release.notes/gb111.release.notes

[25] Distribution Release Notes, Genetic Sequence Data Bank release 167.0, 15 August 2008, ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

[26] NCBI Software Developer's Toolkit, Documentation, http://www.ncbi.nlm.nih.gov/IEB/ToolBox/SDKDOCS/SEQLOC.HTML, downloaded 16 Sep. 2008.

[27] Sample GenBank Record, NCBI, http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html, downloaded 16 Sep. 2008.

Appendix B Accession Numbers

Paraphrasing the NCBI SDK documentation we note the following:

- The three databases (GenBank, EMBL, DDJB) share Accession Identifiers.
- Accessions convey only bookkeeping equivalences between the 3 partners.
- Accessions do not uniquely identify a sequence.
- Accessions may change when only metadata for a sequence changes.
- Accessions may remain unchanged when a sequence is updated.
- There is no shared method of recording the history of a record (accession).
- An Accession is the best identifier to use within the partnership.

Hence, Accession Identifiers identify database records containing metadata about a sequence. If new metadata is received, the Accession Identifier may change. In practice though, the identifier is simply updated with a version number (e.g., AY123456.1 becomes AY123456.2). If only the sequence data changes, the Accession may remain the same. Over time, the base Accession number (typically, though not definitely) remains the same as long as it continues to refer to the same sequenced sample. The Accession with the largest version number will be the current one. Records using earlier Accession versions are not delivered in the database dumps from NCBI.

Appendix C GI Numbers

Again, paraphrasing the NCBI SDK:

- The GI is a simple integer assigned to a sequence by the NCBI identifier (SEQID) database.
- They are assigned for both nucleotide and protein sequences.
- It uniquely identifies a sequence from a particular source.
- If the sequence changes at all (for example, by re-sequencing original material), a new GI is assigned.
- The GI will not change if only metadata is changed.

Hence, the GI uniquely identifies the actual sequence information (and its most current metadata). If the GI number for a record does not change, the sequence is guaranteed to be *character-by-character identical*. Sequences can be identical and be recorded under more than one GI if, for example, they are from different samples or (more rarely) from the same sample sequenced multiple times by different organizations.

Note that GI numbers refer only to the NCBI database. EMBL and DDBJ have separate unique sequence identifier schemes.

Appendix D Summary

Creating stable identifiers from unstable sources is a challenge. The NCBI solution, given in a separate section of the SEQID documentation, offers a careful and specific description of the NCBI solution. A similar discussion appears in the design documentation for the Integrated Gazetteer Database (IGDB). World-wide collaboration on identification schemes is difficult; neither of these two "best" identifiers is perfect.

A succinct summary of the NCBI scheme is this:

- GI numbers give a guaranteed stable reference to a precise sequence of characters derived from a source.
- Accession numbers (sans version) give a fairly stable identifier for a record containing metadata about sequence derived from a source.

A given GI number will always access an identical sequence from the NCBI database. Accession numbers (complete) will always access the same metadata record for a given release of the database. Between releases, the metadata may be updated with or without a corresponding change in the Accession number.