



Genomics for Bioforensics

Marc Colosimo

781-271-7339

mcolosimo@mitre.org

Lynette Hirschman

781-271-7789

lynette@mitre.org

MSR

Problem



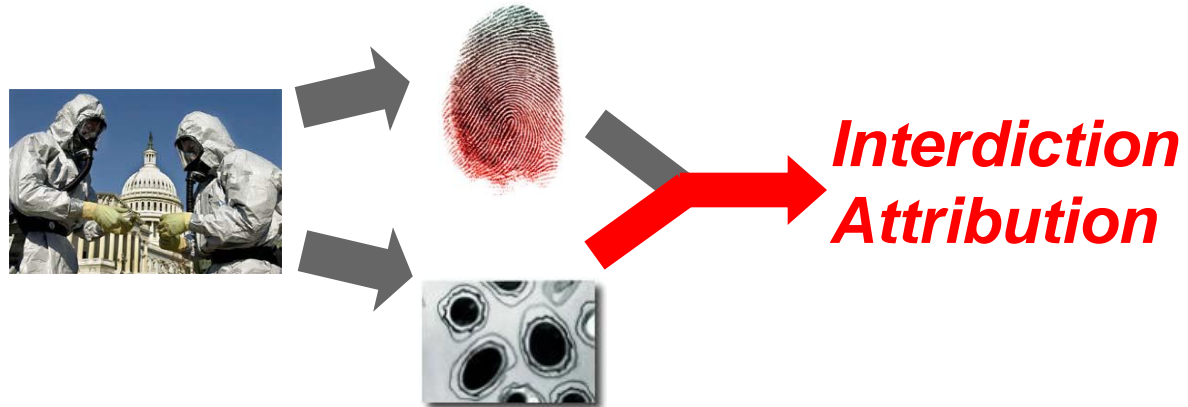
- **How can genomics be used in a biological outbreak for attribution:**
 - To help identify the source of a microbial pathogen or toxin
- **How can genomics be used to distinguish naturally emerging infectious diseases:**
 - From intentionally released pathogens
 - From engineered pathogens

Background

“Classical” Forensics Data

- Fingerprints
- Human DNA fingerprinting
- Trace elements

Sample
Collection



“Microbial” Forensics Data

- Primary identification (pathogen or toxin)
- **Strain identification using genomic information from large sequencing pipelines**

Objective

- **Matching procedure for attribution of biothreat agents**
 - **Clustering**
 - How to automatically create clusters of “similar” sequences
 - How to measure similarities and assign probabilities
 - **Analysis and visualization**
 - How analysts make their decisions
 - What analysts want to see
 - **Reference database (sequences and metadata)**
 - What metadata to collect
 - Where to find data and how to connect different sources (bridging the gap)

Activities

**“I had a little bird,
Its name was Enza.
I opened the window,
And in-flew-enza.”**

- *American Skipping Rhyme, circa 1918*

■ Clustering

- Applying complete composition vectors and affinity propagation clustering to influenza samples

■ Analysis and visualization

- Developing an end-to-end Web-based system that can rapidly be adapted for new types of data

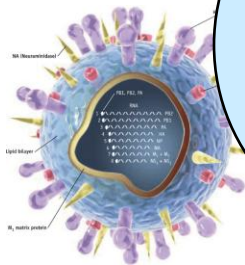
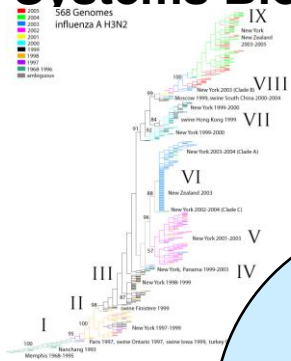
■ Reference database (sequences and metadata)

- Developing InfluenzO, an Influenza Ontology

Highlight

Bridging the Gap—Connecting Genomics and Epidemiology With the Influenza Ontology (InfluenzO)

**Genomic
Sequence Data
Systems Biology**



Genomics:
Genes of
Pathogen

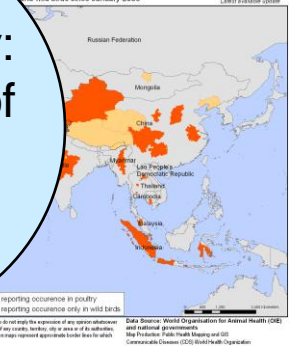
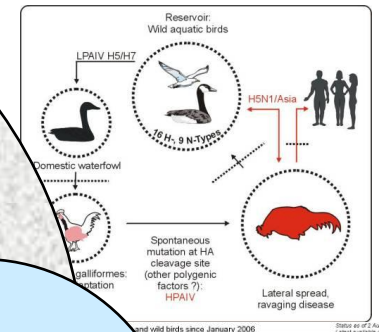
**Geospatial Data
Temporal Data
Pathogenicity
Host**

MITRE

BioHealthBase
BioDefense & Public Health Database

Epidemiology:
Occurrence of
Disease in
Host

**Demographic Data
Clinical Data**

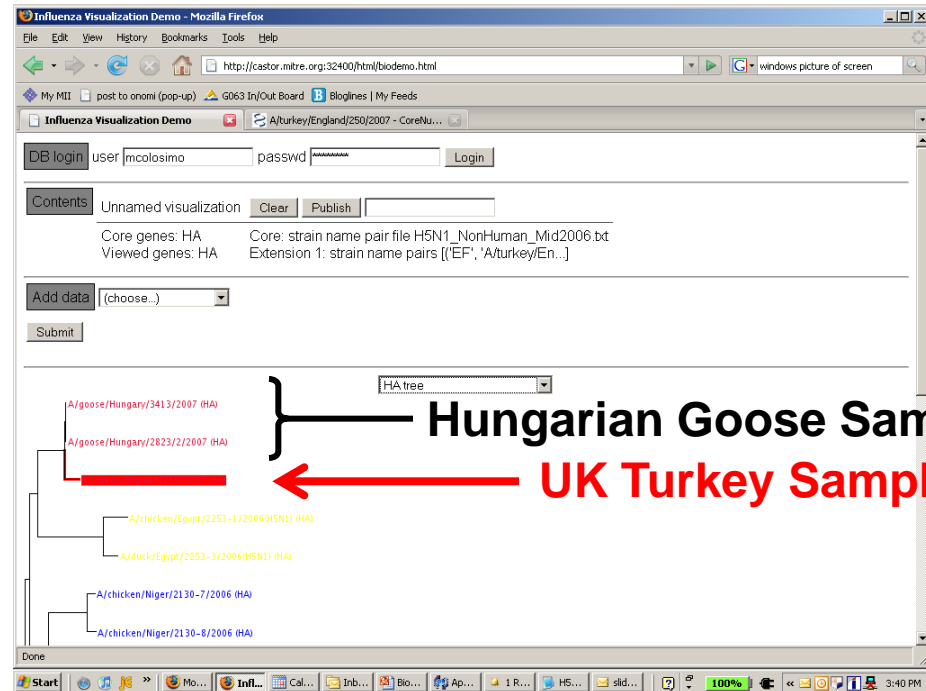


Gemina: Genomic Metadata for Infectious Agents

Demonstration

- Outbreak of H5N1 at a turkey farm, Feb 1, 2007, in the UK
- What is the source of the outbreak
 - Official conclusion: Infected Hungarian poultry was source of H5N1 infection

Closest match in background set of H5N1 over the past 6 months is Hungarian Samples



Impacts

- **The development of algorithms and analytical tools to assist investigators of a bioevent that provides information about relatedness among samples is important for our sponsors**
 - **Complete composition vectors and affinity propagation provide relatedness**
 - **The end-to-end system provides an analytic front end to our tools and data**
- **The development of InfluenzO will provide a unique resource for our sponsors**
 - **We are bridging several communities in forming this ontology**

Future Plans

- **Complete composition vectors/affinity propagation**
 - Validate and refine clustering
- **System integration**
 - Historical data (sequence, epi, etc.)
 - Multiple views of data for analyst
 - Transition to sponsors
- **Ontology development**
 - Complete formalization process and annotate data

