LLNL-TR-452911

# The Microbial DNA Index System (MiDIS): A tool for microbial pathogen source identification

S. P. Velsko

September 2, 2010

## Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

# The Microbial DNA Index System (MiDIS):
# A tool for microbial pathogen source identification

Stephan P. Velsko
Global Security Directorate, S Program
Lawrence Livermore National Laboratory
May 15, 2010

GlobalSecurity Anticipate • Innovate • Deliver

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

# Summary

The microbial DNA Index System (MiDIS) is a concept for a microbial forensic database and investigative decision support system that can be used to help investigators identify the sources of microbial agents that have been used in a criminal or terrorist incident. The heart of the proposed system is a rigorous method for calculating source probabilities by using certain fundamental sampling distributions associated with the propagation and mutation of microbes on disease transmission networks. This formalism has a close relationship to mitochondrial and Y-chromosomal human DNA forensics, and the proposed decision support system is somewhat analogous to the CODIS and SWGDAM mtDNA databases. The MiDIS concept does not involve the use of opportunistic collections of microbial isolates and phylogenetic tree building as a basis for inference.

A staged approach can be used to build MiDIS as an enduring capability, beginning with a pilot demonstration program that must meet user expectations for performance and validation before evolving into a continuing effort. Because MiDIS requires input from a a broad array of expertise including outbreak surveillance, field microbial isolate collection, microbial genome sequencing, disease transmission networks, and laboratory mutation rate studies, it will be necessary to assemble a national multi-laboratory team to develop such a system. The MiDIS effort would lend direction and focus to the national microbial genetics research program for microbial forensics, and would provide an appropriate forensic framework for interfacing to future national and international disease surveillance efforts.

## A CODIS-like system for microbial source inference is needed, and has a sound technical foundation

The National Research and Development Strategy for Microbial Forensics specifies the need to develop a "genetic toolbox for microbial forensics, which addresses the unique requirements of forensic genetic comparisons[1]." Among these requirements are "forensic interpretation guidelines that define what is meant by a genetic 'match' when comparing the genetic sequences of microbial samples[2]" and "… networks and models to help investigators draw inferences regarding sample relatedness with prescribed confidence intervals[3]." Some members of the microbial forensics community have envisioned the creation of a unified system modeled loosely on the Combined DNA Index System (CODIS) and similar human DNA databases to help compare the genetic profiles of forensic pathogen samples to those of reference isolates[4,5]. Until recently the major impediment to the development of such a system was the absence of a sound population genetic framework for calculating association probabilities for microbial isolates analogous to "random match" probabilities in human DNA. However, within the last two years a statistically rigorous framework has emerged[6-8], making it possible to formulate a sensible approach to a microbial genetic index system that can support future microbial forensics investigations.

The key to making probabilistic statements about the relationship between microbial isolates is the observation that most microbial diseases of concern spread and evolve on host-host transmission networks[6,7]. The relevant microbial "population" for answering many forensic questions regarding the origin of an attack strain is the set of all sub-populations of that microbe contained within the nodes of the transmission network, which might include infected hosts in particular outbreaks, or isolates stored in laboratories. The most fundamental type of node is an infected animal or human, although networks of individual hosts can be re-scaled to define more complex nodes such as herds, outbreaks or foci where the more complex node is itself a network of individual infected hosts. New isolates that are created by laboratory passage and exchange of strains between laboratories are also considered nodes in the network. The nodes in the global transmission network for an infectious disease are the potential sources from which a terrorist or criminal would obtain an isolate for nefarious use.

When the genetic sequence of a case isolate is compared to that of a reference isolate from a potential source, we can rigorously formulate and answer questions about the probability that the nodes they came from have particular network relationships. These calculations rely on certain fundamental probability distributions associated with the network and the process of genetic change on the network[6]. Isolates are genetically similar because they were sampled from nodes that were separated by a small number of transmission steps in the network. (Conversely, the probability that the consensus sequences of two isolates differ increases as their network distance increases.) This framework for microbial genetic comparisons is called the "inference-on-networks" theory, and has a close relationship with the forensic analysis of human mitochondrial or Y-chromosomal DNA[7].

The statistics of microbial genetic change among nodes in disease transmission networks provide a consistent framework for formulating and testing forensic hypotheses about strain origin. Examples of the kinds of probabilities that can be estimated within this framework are:

- the probability that a victim was infected by direct transmission from a suspected source (as in cases involving criminal transmission of HIV or other diseases[8].)

- the probability of obtaining a match (sequence identity) between a given isolate and an isolate obtained from a different, randomly sampled node in the same transmission network. This provides an exact definition of how "rare" a particular genotype is in nature[7].

- the probability that two isolates were drawn from the *same* fundamental node i.e. when there is no transmission or laboratory colony selection event separating them. This hypothesis played an important role in the Amerithrax case[9], and is applicable to scenarios where a single batch of agent is manufactured, and then split into sub-batches that are then recovered individually as evidence.

- the probability that an isolate was obtained from a particular past outbreak or disease focus (the "outbreak inclusion probability"). In an investigation of a suspicious outbreak, this can be used to "track" the origin of the attack strain to its ultimate source in nature[6,7].

It is important to recognize that these probabilities cannot be estimated from phylogenetic tree construction alone because at a fundamental level phylogenetic construction ignores the effect that the transmission network has on the microbial genetic population structure[6,7]. The only inference that can be drawn from a phylogenetic tree is that two isolates are genetically more similar to each other than to the other isolates included in the tree. Therefore, concepts of a microbial forensic database that are founded on the notion of a collection of reference sequences that are compared to a case sample using phylogenetics are inadequate and can easily lead to dangerously unsupported inferences about strain origin. The inference-on-networks framework avoids this danger by providing direct estimates of the probabilities of well-defined relationships among isolates.

## A technical vision for a Microbial DNA Index System has been developed

A database and investigation support system for microbial forensics that uses the inference-on-networks framework would be somewhat analogous to the SWGDAM mtDNA Population Database[10] or US Y-STR database[11]. We can designate such a system the Microbial DNA[12] Index System (MiDIS), which would function in a manner analogous to CODIS[13] by helping to narrow and prioritize a set of potential sources for a case isolate in order to generate investigative leads. However, the method for calculating microbial source probabilities is very different from human DNA matching probabilities since neither the "counting method" nor the "product rule" are appropriate. Instead, the microbial database must generate estimates of two basic sampling distributions, designated $P(\delta|M)$ and $\mathcal{P}(M)$, from empirically anchored models that play a role in MiDIS that is analogous to the role played by the population genetics models for human DNA in CODIS calculations.

$P(\delta|M)$ is the probability that two microbial isolates separated by M transmission events will exhibit a genetic difference $\delta$. This is a function of the mutation rate of the pathogen and also depends on the mode of transmission. $\mathcal{P}(M)$ is the probability that two nodes chosen randomly from a transmission network will be separated by M transmission steps. This quantity depends on the size and topology of the transmission network. There are several means for determining these quantities from experimental data drawn from known outbreaks or from carefully designed laboratory studies[6-8]. All probabilistic hypothesis tests for microbial relationships can be computed from $P(\delta|M)$ and $\mathcal{P}(M)$. The core of MiDIS is a computational engine for generating these probability distributions using empirical data, and using them to compute the likelihoods of hypothesized relationships among questioned and reference strains. An architecture for this "microbial inference engine" has been described in detail in a separate document[14].

A useful microbial forensics database that provides parameters to the microbial inference engine would consist of data about known outbreaks of disease associated with each pathogen of concern, including one or more representative sequences from each outbreak, an estimate of the size of the outbreak, laboratory or field data about mutation rates, and certain statistical data about the natural mode of transmission. Depending on the agent, information about human-human, animal-animal, and mixed transmission networks might be required. Networks of laboratories that share strains can be treated using the same formalism.

In any case investigation that involves a pathogen already contained in MiDIS the only input information required from the investigator is genetic sequence data from one or more case isolates and the sequences of isolates from suspected sources. If the hypothesis under consideration is that the attack agent was originally obtained from a known outbreak, it may be necessary to provide an estimate of the network size (number of infected hosts) associated with the outbreak. Previous work[6,8] indicates that calculations of source probabilities are not very sensitive to these estimates.  If a crime or terrorism incident involves a pathogen that is not already contained in MiDIS, then the system provides clear guidance on what reference information must be collected in order to perform hypothesis tests relevant to the investigation. It is also possible that the transmission and mutation properties of the new pathogen can be extrapolated from those of similar, known pathogens.

## MiDIS would provide direction and focus for Microbial Forensics R&D

The development of MiDIS is closely aligned with the goals of the National R&D Strategy for microbial forensics. A major requirement set by the National R&D Strategy is that "…a miocrobial forensic capability must be able to address the requirement to conduct comparative sample analyses in order to query known and questioned samples and draw inferences relating to …. the provenance of a sample or relatedness between samples." MiDIS not only satisfies this requirement, but also would "engage experts in microbial ecology, epidemiology, and agriculture" towards a coherent goal.

Because the inference-on-networks approach integrates information on disease transmission and genetic change, the development of MiDIS necessarily brings together input from many areas:

- Domestic and International outbreak surveillance efforts
- Disease transmission network modeling
- Systematic collection and genetic sequencing of outbreak strains
- Sequencing and typing standards and quality control
- Laboratory and field mutation rate studies

As such, MiDIS becomes a unifying motivation for a requirements-driven R&D program in genetic microbial forensics that can transform the current collection of individual molecular typing and microbial population genetics projects into a coordinated effort that produces a true enduring national capability.

It would be natural to structure the MiDIS program into an initial pilot demonstration phase restricted to a few pathogens, followed by rigorous user-community vetting and validation exercises. Pending a consensus on the utility of such a system, a longer term sustained effort that is concerned with updating and expanding the underlying database to include a larger set of agents, and possibly improving the performance of the microbial inference engine might be initiated. A major goal of the pilot phase is to construct a convincing demonstration of the effectiveness and practical utility of the microbial inference engine concept for microbial agents. Among the microbial agents of interest, *B. anthracis*, *E. coli* O157/H7, and FMD virus have the most extensive pre-existing data sets on which to base the approach and represent "low hanging fruit"[14]. This choice also has the advantage that the host transmission networks for these three diseases are substantially similar, involving primarily cattle or other farm animals, with occasional spillover into wildlife or humans. Therefore, it is reasonable to propose a pilot program for implementation and validation of an inference-on-networks based microbial database and investigation support system focusing on demonstrating this capability for FMDV, *E. coli* O157/H7, and *B. anthracis*.

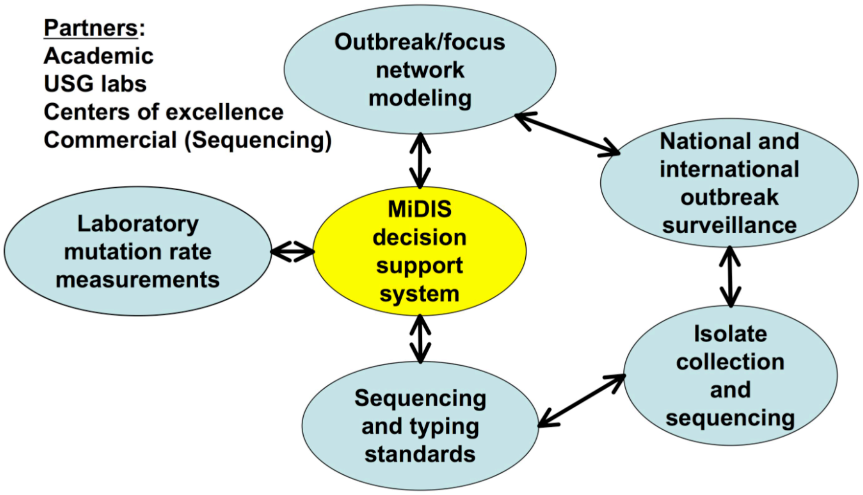
Partners:
Academic
USG labs
Centers of excellence
Commercial (Sequencing)
Outbreak/focus network modeling
National and international outbreak surveillance
Laboratory mutation rate measurements
MiDIS decision support system
Isolate collection and sequencing
Sequencing and typing standards

Figure 1. MiDIS is a natural focus for coordinating and organizing the collection of microbial population genetic data.

## We should begin building the foundations for a sustained community effort to implement MiDIS

It is important to recognize that implementation of MiDIS for a wide variety of microbial agents of concern would require a multi-year, multi-laboratory effort involving interagency coordination and commitment of significant resources over time. Beyond the development of the computational infrastructure and database, the more general effort would involve laboratory experiments, field experiments and epizoological or epidemiological studies, and systematic sequencing of reference isolates. Collaborating organizations would include academic and government laboratories, centers-of-excellence, and commercial institutions. Prior to any such undertaking, it is clearly necessary to build a much firmer organizational foundation than now exists. In particular, it is important to:

1) Develop a prioritized data acquisition/validation agenda
2) Develop a set of applicable standards for data collection and storage
3) Identify potential collaborating laboratories with the required capabilities across the government, academic, and private sectors.

4) Promote transparent peer and user review to establish a consensus on reliability and validity of MiDIS predictions

Developing a pilot demonstration capability for a small set of specific pathogens can be seen as part of a core strategy to accomplish these four goals. First, because the existing foundational data used in the pilot MiDIS program will necessarily be sparse it will naturally lead to a consideration of how to prioritize the acquisition of additional data, which will form the basis of an orderly research agenda. Second, a pilot program provides a tangible framework for drafting and evaluating data standards. Third, to execute the pilot phase itself it will be necessary to engage a team of national collaborators with the requisite expertise in laboratory infection/mutation rate studies, whole genome sequencing, field epidemiology, disease transmission networks, and other inputs that are needed for MiDIS. Finally, the development and validation process will generate specific technical results for peer review, increasing the technical confidence in the underlying approach as well as providing a concrete demonstration of MiDIS capabilities to potential users.

**Notes and references**

1. *National Research and Development Strategy for Microbial Forensics* section I.D.2

2. Ibid, section I.C.1.

3. Ibid, section I.D.6

4. Budowle B, "Defining a New Forensic Discipline: Microbial Forensics", publication #02-12 of the Laboratory Division of the Federal Bureau of Investigation; see also Budowle B, Johnson MD, Fraser CM, Leighton TL, Murch RS, and Chakraborty R, "Genetic Analysis and Attribution of Microbial Forensics Evidence", Crit. Rev. Microbiol. 2005; **31**: 233-254.

5. *Microbial Forensics*, JASON study report JSR-08-512, May 2009

6. Velsko S., Allen J., Cunningham C. "A Statistical Framework for Microbial Source Attribution" Lawrence Livermore National Laboratory Report LLNL-TR-414337, April 30, 2009.

7. Velsko S. "Bacterial Population Genetics in a Forensic Context: Developing more rigorous methods for source attribution", Lawrence Livermore National Laboratory Report LLNL-TR-420003, October 30, 2009.

8. Osburn, JJ and Velsko SP, "Re-evaluating the Florida dentist case with a new statistical framework", Lawrence Livermore National Laboratory Report LLNL-AR-416283, August 28, 2009.

9. The United States Department of Justice Amerithrax Investigative Summary, February 19, 2010. Available from www.justice.gov/amerithrax/

10. Monson KL, et. al., "The mtDNA Population Database: An Integrated Software and Database Resource for Forensic Comparison" Forensic Science Communications, April 2002.

11. Sinha S, et. al. "Utility of the Y-STR Typing Systems Y-PLEX-6 and Y-PLEX-5 in Forensic Casework and 11 Y-STR Haplotype database for Three Major Population Groups in the United States", Journal of Forensic Sciences 2004; 49:691-700.

12. RNA viruses are, of course, also included in this system, although their sequences actually may be stored and compared as DNA sequences.

13. Baechtel FS, et. al., "Tracking the violent criminal offender through DNA typing profiles – a national database system concept", EXS 1991; 58:356-60.

14. Velsko, SP, "Implementing the microbial inference engine for *B. anthracis*, *E. coli,* and FMD Virus", LLNL white paper, July 15, 2010.

15. *National Research and Development Strategy for Microbial Forensics* Goal I.

16. ibid., section I.C.2.