Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

Advanced Statistical Network Models for the Integration of Experimental and Open Source Textual Data for Bioforensic Analyses

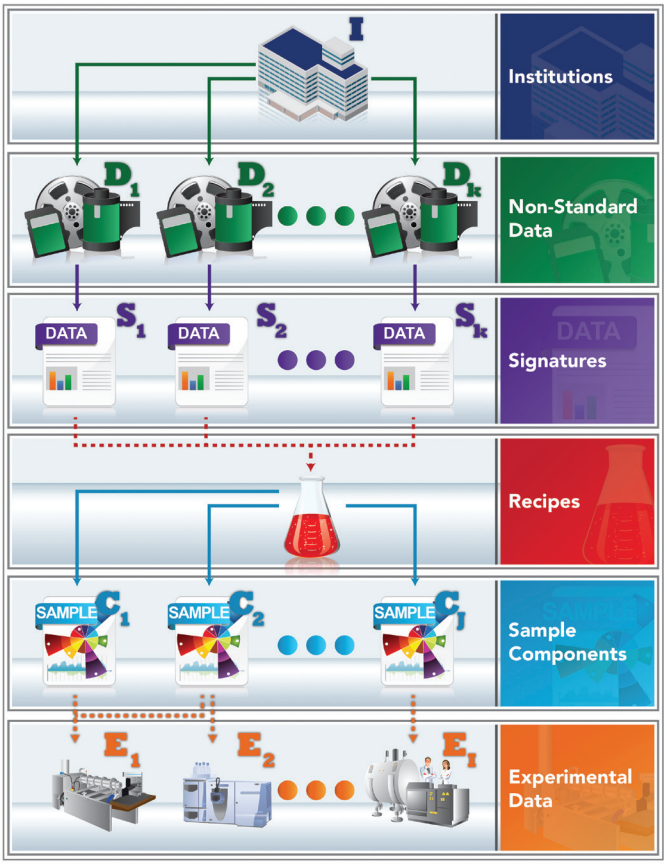
Narrowing the potential source of a biological sample is a key challenge in the forensics sciences because it requires the integration of both hard (e.g., laboratory) and soft (e.g., textual) data in a form that is interpretable by investigators.

CHALLENGE

Biological forensic programs are focused on acquisition of laboratory measurements to identify characteristics of the threat agent (for example, how it was grown). However, to solve a crime, investigators need additional information such as individuals or locations that have the capability to grow the biological agent. This type of information is difficult to generate. In order to make a connection between the forensic signatures and potential source of the agent, it is necessary to consider who could have had access to the relevant source material, equipment, and training to produce the agent in the forensic sample at hand.

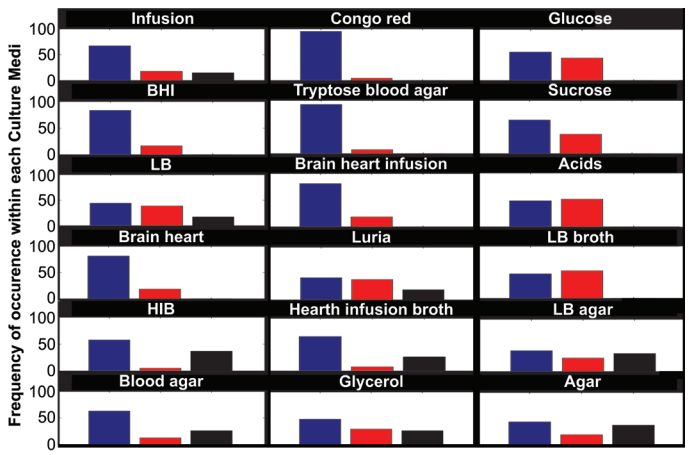
APPROACH/METHODS

Our general approach uses a Bayesian network to integrate experimental and textual data to assign a probability to an institution as having the capability to grow a specific sample. The premise of our model is that scientists at these institutions make public, often through, formal publication their routine culturing practices, such as the types of materials they work with and how they use them. The physical and genetic characteristics of evidence recovered from a biocrime can be compared to these published research practices to point to institutions or research groups that use similar methods.

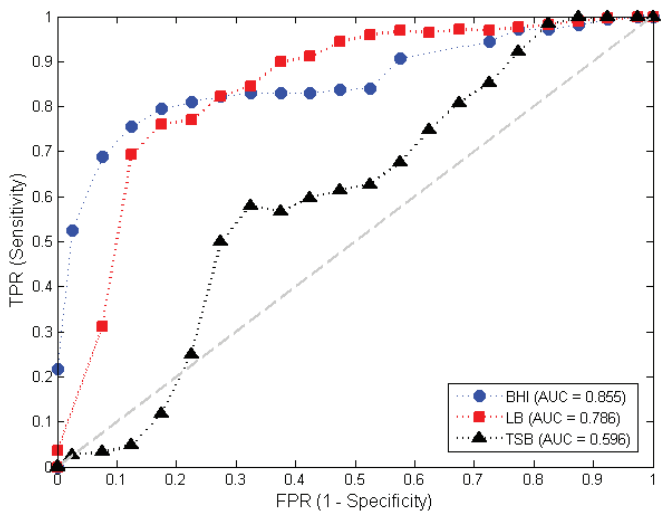


The Generic Bayesian Network to Link Experimental Data from a Sample to Specific End Points of Forensic Interest through Integration with Non-standard Formats of Data (e.g., text)

To date, the fusion of mass spectral data (E) collected on spores on biological agents and textual data in the scientific literature is formulated as a predictive method to obtain the probability that a specific institution (I) could produce a specific sample given the data *(E)*; $P/I \mid E$). The Bayesian network is framed as layers of conditionally independent levels of information that allows the direct prediction of institution from experimental evidence where traditional classification-based approaches would not work (Fig 1). The top and bottom of the network represent the institutions and experimental data, respectively. Since there is no direct link through experimental evidence to institution, we drive the link through the introduction of additional variables that have direct relationships to one



Bar Graph of the Top 18 Key Words Show Various Discrimination of the Three Culture Media



Receiver Operating Characteristic Curves Associated with Each Culture Medium Based on a Model for Yersinia Pestis that Includes 12 Keywords and 5 Experimental Measurements

another. Directly above the experimental evidence in the network are the media components identified through traditional bioforensic analyses. These components are combined to into a recipe. At this point, textual data is related to recipes through numerical descriptors of documents generated through keyword searches. The documents are linked by keywords to institutions.

We have demonstrated the Bayesian framework on several microbial forensics problems including identifying institutions that have the capability to produce the example agent (Bacillus anthracis spores and Yersinia pestis). Fig 2 demonstrates the specificity of certain key words to media recipes and Fig. 3 shows an overall high level of accuracy for the prediction of institution based on the Y. pestis example.

IMPACT/BENEFIT

The primary benefit of the Bayesian integration strategy for hard and soft data is the ability to probabilistically rank institutions for investigators. To our knowledge, no other methodologies currently exist. The longer the lapse between the start of an investigation and the collection of samples for comparison, the more likely the samples within the laboratory will change, decreasing the confidence of matching the samples. Another advantage is that although none of the institutions in the network might currently house the perpetrator, the researchers at these institutions may have valuable information to share with the investigators.

Bobbie-Jo Webb-Robertson
Pacific Northwest National Laboratory
bobbie-jo.webb-robertson@pnnl.gov | (509) 375-2292

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*
U.S. DEPARTMENT OF ENERGY